Energy landscape of the designed protein Top7

Sridhar Neelamraju,*,†,‡ Shachi Gosavi,*,† and David J. Wales*,‡

†*Simons Centre for the Study of Living Machines, National Centre for Biological Sciences, Tata Institute of Fundamental Research*

‡*University of Cambridge, University Chemical Laboratories, Lensfield Road, Cambridge CB2 1EW, U.K*

E-mail: sridharn@ncbs.res.in; shachi@ncbs.res.in; dw34@cam.ac.uk

Abstract

To fold on biologically relevant timescales, proteins have evolved funnelled energy landscapes with minimal energetic trapping. However, the polymeric nature of proteins and the spatial arrangement of secondary structural elements can create topological traps and slow folding. It is challenging to identify, visualise and quantify such topological trapping. Designed proteins have not had the benefit of evolution and it has been hypothesised that de novo designed protein topologies may therefore feature more topological trapping. Structure-based models (SBMs) are inherently funnelled, removing most energetic trapping, and can thus be used to isolate the effect of protein topology on the landscape. Here, we compare Top7, a designed protein with a topology unknown in nature, to S6, a naturally occurring ribosomal protein of similar size and topology. Possible kinetic traps and the energetic barriers separating them from the native state are elucidated. We find that even with an SBM, the potential energy landscap (PEL) of the designed protein is more frustrated than that of the natural protein. We then quantify the effect of adding non-native hydrophobic interactions and coarse-grained side-chains through a frustration density parameter. A clear increase in frustration is

observed on including side-chains, whereas adding hydrophobic interactions leads to a narrowing of the funnel and a decrease in complexity. The most likely (un)folding routes for all models are derived through the construction of probability contact maps. The ability to quantitatively understand and optimise the organisation of the PEL for designed proteins may enable us to design structure-seeking landscapes, mimicking the effect of evolution.

Introduction

The energy landscape of a protein determines its structure, dynamics, and thermodynamics.[1,2] A consequence of the energy landscape theory[2–4] and the principle of minimal frustration[5] is that most natural single-domain globular proteins have evolved a funnelled energy landscape to fold robustly to the native state with folding rates that are biologically feasible.[6] A natural protein folds to its native state through progressive evolution of an ensemble of native-like partially folded structures. This folding mechanism is controlled by both the shape of the landscape and the degree of frustration. Here, frustration is defined in terms of low-lying minima with distinct partially folded structures, separated by high barriers.[7,8]

We employ discrete path sampling (DPS)[9,10] to construct a database of connected local minima and transition states on the potential energy landscape (PEL), which is visualised using disconnectivity graphs.[11,12] DPS has been employed to sample the PEL of several biological systems,[13–18] including a 16 residue β-hairpin peptide[16] and an RNA tetra-loop hairpin.[19] More recently, DPS was used to evaluate the energy landscape of an intrinsically disordered protein,[20] and a coiled-coil conformational switch.[21] Further, a three-colour (BLN) 69-residue model protein was constructed and shown to be fully funnelled in its Gō model representation.[22] Gō, or structure-based models (SBMs), encode the native structure of the protein, usually leading to funnelled energy landscapes, which isolate the effects of topology by removing energetic trapping due to non-native interactions.[23] SBMs reproduce correct folding trajectories for a diverse set of natural, single-domain, globular proteins in

simulations.[6,24]

It has been postulated that designed proteins are more frustrated than their natural counterparts.[25,26] Top7 was designed specifically to exist in a topology unknown in nature.[27] It has been hypothesised that some of the frustration present in its energy landscape is due to this non-natural topology.[28] The thermodynamics and kinetics of Top7 have been extensively studied with SBM-like models,[29–31] thus providing useful benchmarks for assessing the applicability of DPS to examine frustration in designed proteins. Previous theoretical studies on folding[24,29–31] have concluded, on the basis of free energy profiles derived from molecular dynamics with native-centric SBM-like potentials, that multiple metastable states exist in Top7 with the C-terminus preferentially formed, while the N-terminus remains unfolded. This picture is confirmed through experiment[28,32,33] where the C-terminus forms a stable homo-dimer, with a secondary structure that remains nearly unchanged up to 98° C and with high concentrations of denaturant.[32] In comparison, the free energy profiles derived for evolved natural proteins of a similar size (like S6[34]) indicate that these mostly fold in a cooperative two-state manner.[30,31] Watters et al.[33] found that below 4 M guanidine concentrations, the folding of Top7 is a triple-exponential process, indicating that at least four states are populated during the folding reaction. A "cavity" was reported in the N-terminus half of Top7 and modifications to this region were recently suggested to make it fold more cooperatively.[35] Further, Zhang et al. noted that non-native hydrophobic interactions are essential for the roll-over observed in the folding arm of Top7 in the chevron plot.[30] Here, the hydrophobic residues V48, F63, A64, A65, L67, and V81 were implicated in stabilising non-native states of Top7.

The naturally occurring ribosomal protein S6,[34] is similar to Top7 in size and shares with it many topological features. The contrasting folding behaviour of the two proteins has been studied in depth[24,31] with native-like SBM models and molecular dynamics. Further, a tetra-peptide fragment-based design algorithm was used to design another 92 residue protein, M7,[36] which folds into a structure identical to Top7.[37] M7 was designed specifically to fold

more like a natural protein, i.e. in a two-state manner. Thus, it should be insightful to construct energy landscapes for these three proteins.

We define a frustration density parameter[38] that is able to quantitatively capture the frustration of the PEL at low energy for the three proteins. Change in frustration densities over a large part of the (un)folding route are derived and compared for Top7, S6 and M7. The preferred paths for (un)folding of Top7 and M7 are also compared. Further, using Top7 as an example, we show that DPS, in combination with SBMs, is a viable tool to find low-energy intermediates and quantify frustration in the PEL of a protein.

Methods

Potential energy functions

In a Cα-SBM, each residue in a protein is coarse-grained to a spherical neutral bead centred at the α-carbon. The native topology is defined purely on the basis of a contact map derived from the native state, which is then projected onto the coarse-grained Cα representation. Thus the number of native contacts (with identical interactions) is predefined. Each residue in contact with another in the contact map has an attractive Lennard-Jones interaction. In principle, this contribution biases the energy landscape in favour of the native state, while eliminating kinetic traps emerging from energetic frustration (accounted for through non-native interactions) yielding a funnelled potential energy landscape.

Contact maps are generated with all heavy atoms and then projected onto the Cα representation. For Top7, contact maps for the Cα-model are generated in two ways; first, with the shadow contact map (SCM) algorithm,[8,39] where parameters are derived from the SMOG server[39] in the Cα representation with a cut-off radius of 6.5 Å, and second, with a cut-off based contact map where the cut-off radius is set at 4.5 Å[35] (see figure 1). The size of each individual Cα bead is set at 4 Å as was the case in previous studies.[24,29–31] When we want to introduce solvent effects in the Cα-model, we allow beads belonging to non-native

hydrophobic residues (A, V, L, I, M, F, W and Y) to interact with a 12-10 Lennard-Jones type potential, again with an interaction strength of ϵ and a constant distance separation of 5.5 Å.[40] The intramolecular effect of water molecules is represented by a double-well desolvation barrier potential.[41,42] Here, in keeping with previous studies,[24] the 4.5 Å cut-off based contact map was used for native-contacts.

The overall SBM employs two stiff harmonic terms that restrain the bonds ($K_b = 100\epsilon$) and angles ($K_\theta = 40\epsilon$) among the beads, a dihedral term ($K_\phi = 1\epsilon$), and an attractive Lennard-Jones term, which defines attractive interactions among beads that are found to be in contact from the contact map. A repulsive excluded volume term, which defines the short-range repulsions between the beads that are not in native or non-native contact, is also added. Throughout the manuscript we use reduced reduced energy units with $\epsilon = 1$ kcal/mol.[8,39]

A simple two-bead Cα-Cβ model as defined by Cheung et al.[43] was employed to account for side-chain interactions. A Cβ bead was placed at the centre of mass of the side-chain of each residue. Then the native structure was partitioned into backbone and side-chain atoms, from which Cα-Cα, Cα-Cβ and Cβ-Cβ contacts were derived.[1]

Mapping the energy landscape:

Molecular Dynamics

Starting from the native state, molecular dynamics simulations were performed approximately at the folding temperature (T_f) where the folded and unfolded states have similar populations. The topology files for the SCM map were generated with SMOG.[39] All other topologies for GROMACS and OPTIM were generated with an in-house code. A leapfrog stochastic dynamics integrator was used with a time step of 5 femtoseconds. We checked that several transitions occurred between folded, unfolded and intermediate states in order to ensure extensive sampling. An ensemble of conformational states was derived for S6 and Top7 from these molecular dynamics simulations.

[1]Details of potential parameters are included in Supplementary Information

As seen in earlier studies for Top7,[7,24,29,31,33] an intermediate state is populated at $Q \approx 0.56$. The intermediate state disappears for the ensemble derived with the shadow-contact map with a 6.5 Å radius. The addition of non-native hydrophobic attractive interactions leads to a surge in the number of intermediate states obtained in Top7, while the evolved S6 protein displays very little change in folding behaviour. From this result, earlier studies concluded that naturally occurring proteins of similar size, such as S6, fold significantly more cooperatively than the designed protein, Top7. In Figure 2, we confirm these conclusions for different flavours of SBMs employed in our study. In each case, the free energy profiles for S6 showed cooperative behaviour and a lack of intermediate states. In contrast, the free energy profile of Top7 is quite sensitive to the type of contact map and potential employed. Multiple intermediates appear, depending on the type of perturbation applied. Thus, with the confidence that the potentials employed in this study reproduce known results, disconnectivity graphs were constructed with DPS.

Discrete path sampling (DPS)

For a given protein, an initial set of local minima was generated from conformations derived by clustering the molecular dynamics trajectory into folded, unfolded and intermediate states. Each of these geometries was minimised with a modified L-BFGS algorithm,[44] as implemented in OPTIM,[45] for a convergence threshold on the magnitude of the gradient at 10^{-6} kcal/mol. A database of stationary points (local minima and transition states) was then constructed with discrete path sampling as implemented in PATHSAMPLE,[46] by connecting pairs of local minima. A transition state is defined as a first order saddle point on the potential energy surface, where the Hessian has exactly one negative eigenvalue, whose eigenvector corresponds to a local reaction coordinate. New stationary points are accepted and added to the database on-the-fly as they are found. The missing connection algorithm[47] was used to select pairs of minima. Initial transition state guesses were generated and lowest-energy paths connecting a pair of local minima were found with the doubly-nudged[48,49] elastic band[50,51]

(DNEB) method, followed by interpolation between the end points.[52] Further optimisation was carried out by hybrid eigenvector-following in OPTIM.[53] The network of connected stationary points thus derived for the protein was visualised as a microcanonical disconnectivity graph.[11,12,542]

In this graph, each minimum corresponds to a vertical branch terminating at the corresponding potential energy. For a regular series of threshold energies spaced by a chosen value δ, we perform a superbasin analysis[11] to separate the minima into disjoint sets, whose members can interconvert without exceeding a transition state energy above the current threshold. The branches join at nodes on the vertical (energy) axis as the superbasins merge. The arrangement on the horizontal axis is usually chosen to ensure that the branches do not cross, with the lowest-lying minima in the centre. The size of the energy spacing parameter δ is chosen to visualise the structure of the landscape in an informative way. Further details can be found in Ref.[12,14,55]

For de novo design of a minimally frustrated protein, it is useful to have a metric that quantifies frustration, which can be used to optimise a given sequence through mutation. "Z-score" and "double Z-score" statistics , derived from the random energy model,[56] are based on the ratio of folding temperature (T_f), to the glass transition temperature (T_g), and have been suggested as metrics to quantify frustration.[57,58] In this model, T_f/T_g is proportional to the ratio of the energy gap between the folded and unfolded states ($\delta E_s = E_u - E_f$), and the fluctuations in energy in the unfolded ensemble ($\Delta E = \sqrt{(E_f - \langle E_f \rangle)^2}$). δE represents the height of the folding funnel, and ΔE the frustration. Thus, $T_f/T_g > 1$ essentially implies a globally funnelled landscape, which can be visualised by a disconnectivity graph. A fully funnelled landscape is characterised by a "palm tree" appearance.[12]

Here, we define frustration density (ρ) as the fraction of minima discarded from the branch with the native state at the bottom at a given energy level, divided by δ, the spacing between energy levels.[38] This metric is directly accessible from the disconnectivity graph.

[2]See supplementary information for input files, potential parameters and details of sampling.

Other metrics for measuring funneling/frustration include the route function of Clementi et al.,[6] based on the probability of forming native contacts, a geometric "topological folding barrier" determined through a Delaunay tesselation,[59] and more recently the frustration parameter of de Souza et al.[60]

Results and discussion

Discrete Path Sampling: Cα-SBMs: Top7, S6 and M7

Construction of a disconnectivity graph requires an appropriate description of reactant and product states. Multiple conformers of Top7 were drawn from the molecular dynamics trajectories (described in Figure 2) and clustered into the unfolded state ($Q < 0.4$), the intermediate state ($0.4 \leq Q < 0.7$) and the folded state ($Q \geq 0.7$). L-BFGS minimisations were then performed to obtain a set of around 5000 local minima, which were used as starting points for constructing the disconnectivity graph.

In experiment it is observed that the C-terminus of Top7 is always preferentially folded.[32,61] The presence of a topological trap at the N-terminus could explain this behaviour to some extent. The C-terminus region remains fully folded in our ensemble up to the threshold studied (i.e. $Q = 0.65$) above the native state. Truong et al.[31] found intermediates for Top7 with a non-additive SBM[62,63] (but at a Q value of around 0.35). The intermediates contained an entirely unfolded N-terminus and partially unfolded C-terminus. A non-additive SBM includes three-body terms in the Hamiltonian to improve prediction of experimentally determined rate constants.

From the disconnectivity graph, the structures labelled A1, A2 and A3 were characterised as possible traps by the appearance of small sub-funnels. We define a sub-funnel (or a kinetic trap) as a branch (other than the main branch associated with the global minimum) containing more than 50 nodes at δ=0.5. This choice allows for easier visualisation and characterises the traps found for the proteins in question. Three different sub-funnels were

observed and the conformers at the bottom of these funnels are shown in Figure 3. The presence of these funnels in a Cα-SBM with identical contact strengths indicates that the topology of the Top7 protein itself might lead to trapping. We note that each of the basins, corresponding to possible kinetic traps, do not appear to occupy a large volume in phase space, and thus may not be easily accessible (from MD).

In A1 the β1-β2 strands are similar to the native state, N, but are shifted in a concerted manner around an elongated α2-helix. This conformer resembles the intermediate structure I_2 reported by Zhang et al.,[30] with a Cα-SBM that includes a desolvation barrier and non-native hydrophobic interactions. The contact map used in that study was identical to our 4.5 Å cut-off based contact map with a total of 201 contacts. The A1 sub-funnel is connected to the A2 basin, as shown in Figure 3. In A3, the β1 and β2 strands break all contacts and then reform them after accommodating the loop between β3 and α1 motifs. While all beads were free to move, it is always the N-terminus that is deformed for up to $55\,\epsilon$ above the native state. The C-terminus of the protein,[27] is fully intact during unfolding in the low-energy region. A cavity on the N-terminus has been previously noted as a probable cause of frustration.[35] Curiously, when the SCM algorithm was used to define the contact map, we observed that the C-terminus of the protein was the most likely to unfold. It is possible that folding routes of designed proteins may be more sensitive to small changes in the contact map. In further analysis, we always use contact maps derived from the 4.5 Å cut-off radius map as it reproduces the experimentally known pathway.

Truong et al.[31] reported that the free energy profile of Top7 is considerably more complex than that of the natural protein S6. In Figure 4, we show disconnectivity graphs for M7 and S6 derived from the Cα-SBM. No sub-funnels (traps) were found in the graphs derived for S6 and M7 that satisfy our criterion of 50 nodes at δ=0.5ϵ. Instead, we see a funnelled "palm-tree" like disconnectivity graph. We discuss the distribution of frustration in the landscapes represented in Figure 3 and Figure 4 in a later section.

Effect of adding non-native hydrophobic attractions on the PEL of Top7

Enhanced structure-based models (eSBMs),[35] i.e. SBMs enhanced with non-native interactions, have been used to study the folding behaviour of many proteins.[24,30,35,40] A multi-dimensional microcanonical disconnectivity graph can facilitate the tuning of enhanced structure-based models over a significant part of the low-energy landscape. The PEL of Top7 using a Cα-eSBM with non-native hydrophobic interactions shows that the energy separation between near-native and native states (a quantity often used to describe frustration) remains the same ($10\,\epsilon$), but a clear narrowing of the funnel is observed (see Figure 5). We note that in each case, the graphs were constructed up to $55\,\epsilon$ (around $Q = 0.65$ for ϵ=1 kcal/mol) above the native state. This result agrees with several molecular dynamics studies, where addition of non-native interactions between hydrophobic contacts leads to an increase in the rate of folding,[64] and frustration appears to decrease.[65]

Structure A4 (Figure 5) was identified at the bottom of the only sub-funnel found in the graph. β1-β3 are able to break contacts and reform them around the loop connecting the α2 and β3 motifs. From the frequency of contact formation in the low-energy ensemble (Figure 7), it becomes clear that hydrophobic contacts also destabilise the β1-β3 interactions.

Effect of adding coarse-grained side-chains on the PEL of Top7

The addition of side-chain side-chain and backbone side-chain interactions to the standard backbone-backbone interactions through a native-centric Cα-Cβ model makes the topology more complex. This change might lead to a more frustrated landscape, with multiple sub-funnels in the disconnectivity graph. We discuss the frustration more quantitatively in a later section. The structures labelled A5, A6, A7 and A8, which lie at the bottom of the four largest sub-funnels found, are depicted in Figure 6, along with the disconnectivity graph of the Cα-Cβ model of Top7. These structures correspond to possible low-energy intermediates

(see Figure 6) that are expected to arise from the Cα-Cβ model. Here, we consider conformers up to 100 ϵ above the native state and extensive sampling was performed.

Interestingly, structure A7 is very similar to the intermediate I_0, suggested by Zhang et al.[30] In I_0, the helical contacts are broken leading to a bending of the connected β-strands 1 and 3. This observation suggests that A7 results from quenching I_0 (minimising with an L-BFGS routine in our case). Two of the reported three intermediate states (I_0, I_1 and I_2[24,30]) were also characterised as sub-funnels in the Cα-Cβ model, possibly emphasising the role of side-chain interactions in the folding behaviour of Top7. Finally, two other intermediates suggested by Truong et al.[31] at $Q = 0.25$ and $Q = 0.35$ have not been sampled in the present work,as they require the entire N-terminus to unfold, along with parts of the C-terminus. Analysis of this unfolding would require extensive sampling to a higher energy threshold above the native state. Use of more elaborate predictive models, such as AWSEM[66] and statistical potentials[26] for a more realistic description of folding processes in natural and designed proteins will be the focus of future work.

Finally, in supplementary Figures S1 and S2, we provide a comparison of the kinetic traps. The potential energy of structures A1-A8 relative to the native state is plotted with five different potential parameters. For structures identified with the two-bead potential (A5-A8), the Cβ atoms were removed to make a comparison with the Cα-only structures. In A5-A8, the effect of adding hydrophobic interactions is minimal, as the energy does not change much with respect to the native state. In structures A3 and A4, adding hydrophobic interactions increases the relative energies of the structures. Relative energies calculated from the two-bead model for A5, A7 and A8 are larger by about 30 ϵ than the Cα-only counterparts. However, in the case of A6, the relative energies derived from the SCM Cα-SBM are similar to the two-bead SBM. This result might indicate that side-chain interactions do not play an important role in stablising this particular structure. Figure S2 depicts the 'fastest' paths found for transition to the native state with four different potential parameters. Here, the effect of adding hydrophobic contacts on the lowest-energy transition path are

apparent. In A1, A2 and A3, a significant increase in barrier height is observed along the transition path when hydrophobic interactions are added.

Preferred mechanism of (un)folding: Top7 with various SBM flavours and M7

We construct "probability contact maps" for each ensemble comprising only the local minima connected to the native state (see Figure 7). Analogous maps are also constructed for the molecular dynamics ensembles. The map derived for the Cα-SBM with a 4.5 Å cut-off (with 201 contacts) is treated as the base on which each of the other contact maps are projected.

A combination of single-molecule force microscopy and steered molecular dynamics simulations with Top7 suggested that, for mechanical unfolding, the backbone hydrogen bonds connecting β strands 1 and 3 break concurrently as the N-terminus and the C-terminus slide past each other.[67] Thermodynamically as well, from Figure 7a this process appears to be the preferred path. All the potentials, with the exception of the 6.5 Å cut-off SCM map,[39] indicate that the β1-β3 contacts are formed with the lowest probability, and that unfolding proceeds from the N-terminus. Addition of non-native attractive hydrophobic terms exacerbates the destabilisation of the contacts between the β1 and β3 strands. Furthermore, contacts among these two strands are also the most likely to break in the Cα-Cβ model, which includes side-chain effects.

The contact map of Top7 derived from the SCM algorithm (Figure 7a) changes the unfolding route such that the C-terminus unfolds first. Previous MD studies[29,31] have noted a change in the folding route, and we can now clearly attribute the change to the unfolding of the C-terminus as we have the benefit of considering only the minima connected to the native state through one or more transition states.

The folding route of a designed protein could be more sensitive to the contact map definition. Sixty additional contacts added from the SCM algorithm (shown in the bottom right half of Figure 1b as blue squares) completely alter the preferred mechanism for unfolding.

Interestingly, the cooperatively folding topological analogue of Top7, the protein M7, also unfolds from the C-terminus first, irrespective of the type of contact map used. Thus, while the overall topologies of the two proteins are identical, we can see a clear preference for unfolding to proceed from the other end for M7. We can also see, with the evidence from the folding behaviour of M7, that the topology of Top7 (and M7) is indeed conducive to cooperative folding.[30] Very little experimental data exists for M7 and to the best of our knowledge, there are no experimental measurements to verify the unfolding mechanism.

Frustration: Top7, M7 and S6

We measure the frustration of the potential energy surface in terms of a density parameter (ρ), defined as the number of (normalised) minima that branch off the main funnel at a given energy level divided by the width of the energy spacing (δ) in the disconnectivity graph. In Figure 8, we compare ρ for the landscapes derived for a) Top7, Cα-SBM (4.5 Å cut-off), b) S6, Cα-SBM , c) Cα-Cβ-SBM and d) M7, Cα-SBM. We note that this analysis is limited to the low-energy parts of the PEL (around $55\,\epsilon$ for Cα and $100\,\epsilon$ for Cα-Cβ models) as most of the sampling was performed under these threshold values in each case. As ρ is a function of the magnitude of δ we plot it for four different values, namely $0.5\,\epsilon$, $1\,\epsilon$, $2\,\epsilon$ and $4\,\epsilon$ (ϵ=1 kcal/mol). Qualitative differences in the frustration become quite evident by tuning δ.

Up to $55\,\epsilon$, ρ for Top7 fluctuates, peaking first around 15ϵ, falling, and then peaking again around $45\,\epsilon$ above the native state. For S6, very little fluctuation is found in ρ. A structure-seeking system is one that can relax to the global minimum rapidly. Examples include magic number clusters, crystals, and evolved single-domain proteins.[1] As is characteristic of structure-seeking landscapes, ρ increases exponentially as we move further away from the native state. The Cα-Cβ model of Top7 exhibits many fluctuations over a small energy range. This result indicates that more pockets of local minima are being separated into sub-funnels. ρ increases continuously up to around $40\,\epsilon$ above the native state, after which

fluctuations begin to appear. Here, ρ peaks between 26 and 35 ϵ above the native state. The first peak can be viewed as an indication of the funnel depth. An analysis of this kind is likely to be useful when determining the energy separation between the native states and near-native states for application to the random energy model. The T_f/T_g parameter is used as a proxy for this energy difference for explicit negative design of cooperatively folding proteins.[31,58] A higher ρ peak indicates a deeper global funnel, and quantitatively determines the difference in energy between near-native and native states.

It was expected that the Cα-Cβ model would lead to greater frustration, but it appears that below 40 ϵ the potential energy surface is quite funnelled; the expected increase in ρ contributed by side-chains begins to appear only above this energy. Here, we are able to show quantitatively that SBMs, which are more funnelled by design, can capture the increase in frustration of the potential energy surface of a protein only designed to fold into its native state, but where the design process has not optimised the folding route. For S6, we can see how the funnel evolves with very little fluctuation in frustration density. Interestingly, for M7 as well, the funnel remains quite well defined up to about 40 ϵ with very little fluctuation in ρ, although not as pronounced as it is in S6 underlining the evolved property of minimally frustrated funnels in natural proteins.

Conclusions

We have investigated the nature of the potential energy landscapes (PELs) for various coarse-grained structure-based models (SBMs) of a designed protein with a non-natural topology, Top7, its sequentially different but structurally similar designed analogue, M7, and the natural protein with similar topology, S6, using discrete path sampling (DPS). The landscapes were visualised using disconnectivity graphs and analysed in a variety of different ways. We find that DPS can identify known kinetic traps along the folding path of Top7, which manifest as sub-funnels in the landscape. Further, despite encoding the structure and having little

energetic frustration, Cα-SBMs capture the increase in frustration in the energy landscape of Top7, when compared to S6 and M7.

By analysing two different landscapes of Top7, which are calculated using Cα-SBMs that encode slightly different contact maps, we find that the two models fold by different routes. Thus, designed proteins may be more sensitive to small changes in the contact map. We then add non-native hydrophobic contacts to the Top7 Cα-SBM and this leads to a PEL with a narrower funnel. In a different perturbation to the model, we increase the complexity of the structural description and add Cβ side-chain beads to the Top7 Cα-SBM. A clear increase in frustration is observed at higher energies with the Cα-Cβ-SBM. The changes in frustration upon changing model or protein are quantifiable with a frustration density parameter. Overall, we conclude that the PELs of designed proteins are likely to be less robust to changes in model and more frustrated than those of natural proteins. DPS can be used to quantitatively understand the PELs of SBMs of both designed and natural proteins. This insight could in future be used to construct a mutational strategy to sculpt structure-seeking funnelled landscales in artificial proteins.

Figures

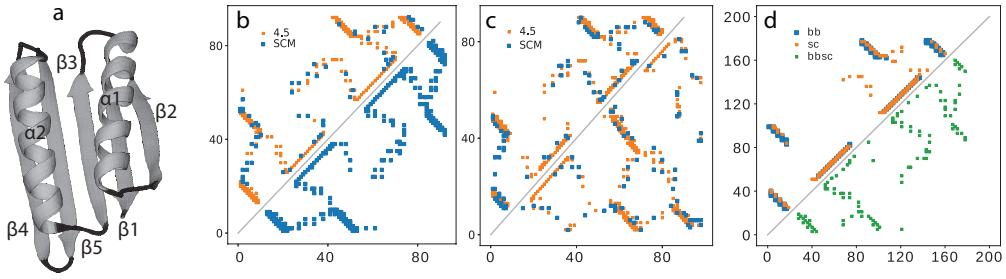

Figure 1: (a) Top7 (PDBID:1QYS) with labelled secondary structure motifs. (b) Contact maps for Top7 derived from the 4.5 Å cut-off map (top-left, orange squares) and the 6.5 Å SCM algorithm[39] (bottom-right, blue squares). Blue squares on the top-left depict the extra contact-pairs added from the SCM algorithm. (c) Contact maps for M7 (PDBID: 2JVF, top-left) and S6 (PDBID: 1RIS, bottom-right) derived from the 4.5 Å cut-off map (orange squares). Blue squares depict contacts present in the SCM map but absent in the 4.5 Å cut-off map. The horizontal and vertical axes represent each bead (Cα or Cβ atom) numbered in ascending order from N- to C-terminus. (d) Contact maps in the α-Cβ representation[43] for Top7 coarse-grained to 179 beads. The top-left half depicts backbone-backbone (bb, blue squares) and side-chain side-chain (sc, orange squares) beads in contact in the native state. The bottom-left half depicts the backbone beads in contact with side-chain beads (bbsc, green squares).

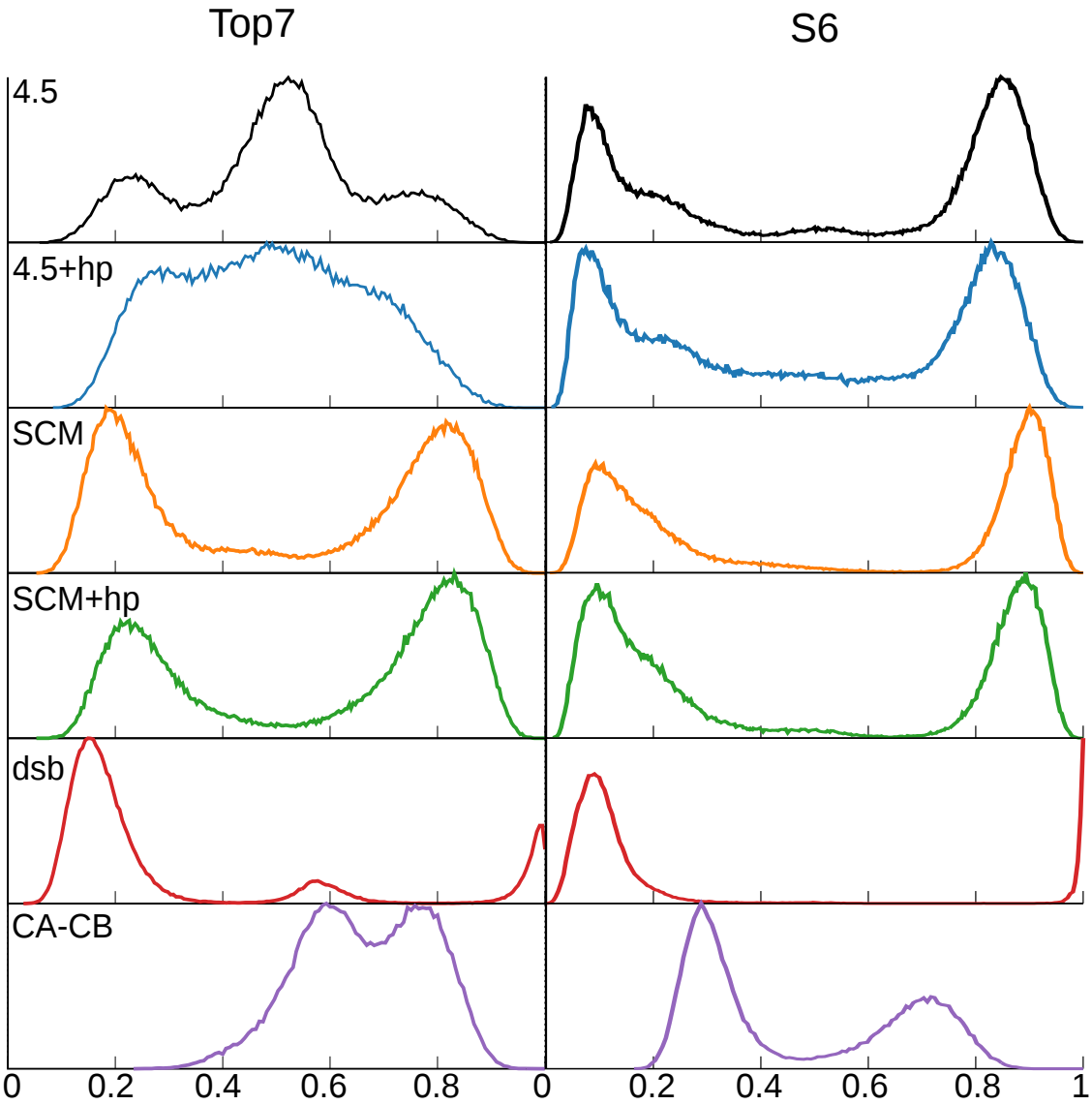

Figure 2: Normalised population of conformational states (vertical-axis) as a function of fraction of native contacts, Q (horizontal axis), from a molecular dynamics trajectory at around T_f for contact maps derived from a 4.5 Å cut-off (4.5, black), a 4.5 Å cut-off with non-native hydrophobic interactions (4.5+hp, blue), a 6.5 Å shadow contact map (SCM, orange), a 6.5 Å shadow contact map with non-native hydrophobic interactions (SCM+hp, green), and a 4.5 Å cut-off map with contacts interacting with a desolvation-barrier type potential (dsb, red) instead of the 10-12 LJ potential, the Cα-Cβ model[43](CA-CB, purple). Clearly separated folded and unfolded states appear in the evolved ribosomal protein S6 (right), whereas intermediate structures appear for the designed protein Top7 (left).

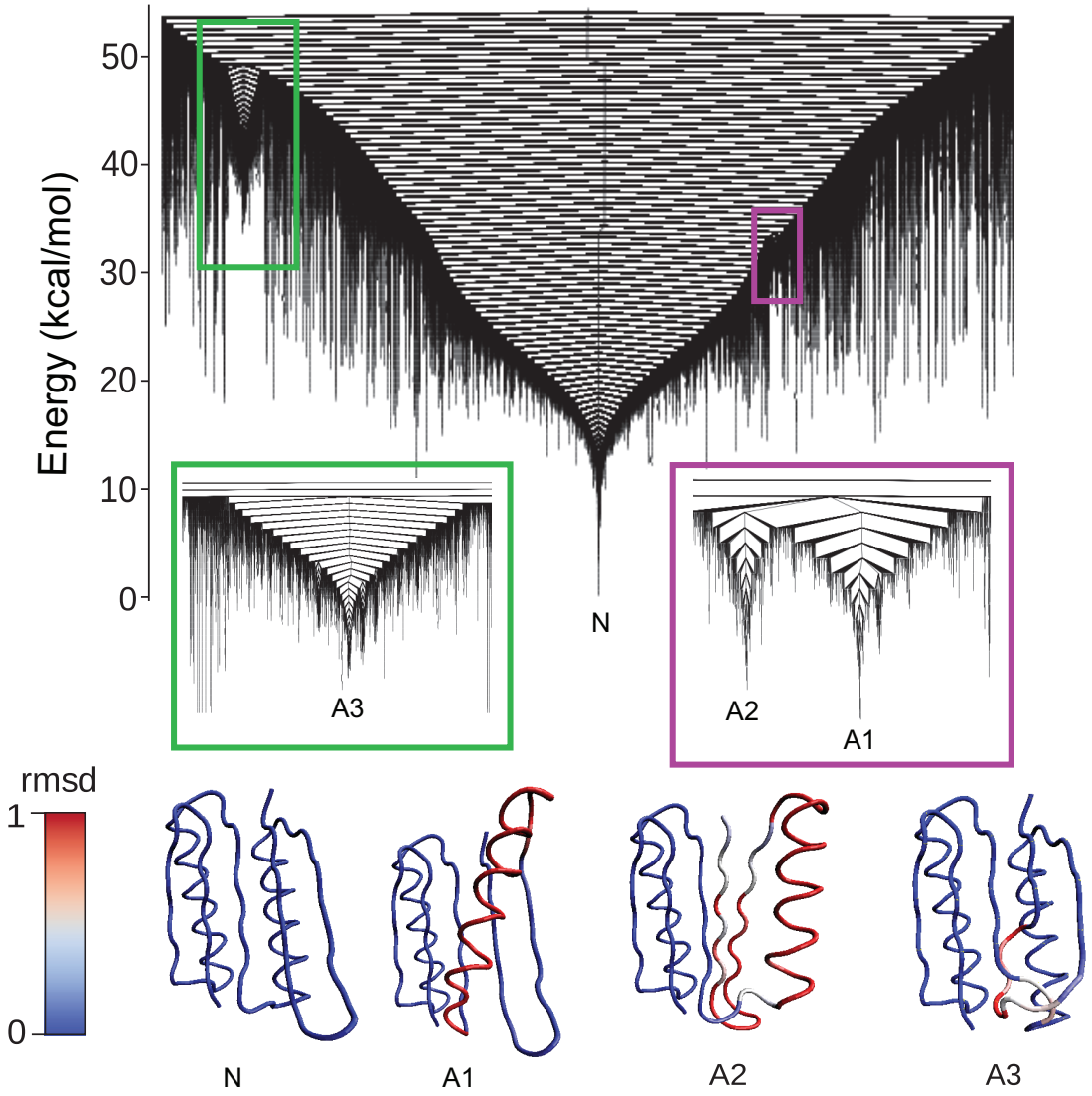

Figure 3: Disconnectivity graph representation of the ensemble of stationary points derived with DPS for the 4.5 Å cut-off contact map. Thorough sampling up to 55 kcal/mol (ϵ) above the native state was performed with 526,166 minima and 611,148 connected transition states. Two sub-funnels with more than 50 nodes are shown in purple and green. The boxes are magnifications. Conformers corresponding to the bottom of each funnel are labelled N (Native state), A1, A2 and A3, conformers corresponding to the bottom of each sub-funnel are depicted at the bottom of the figure. Blue indicates the parts of the protein similar to the native state (N) and red indicates a high RMSD difference obtained from an RMSD-based per residue structural alignment.

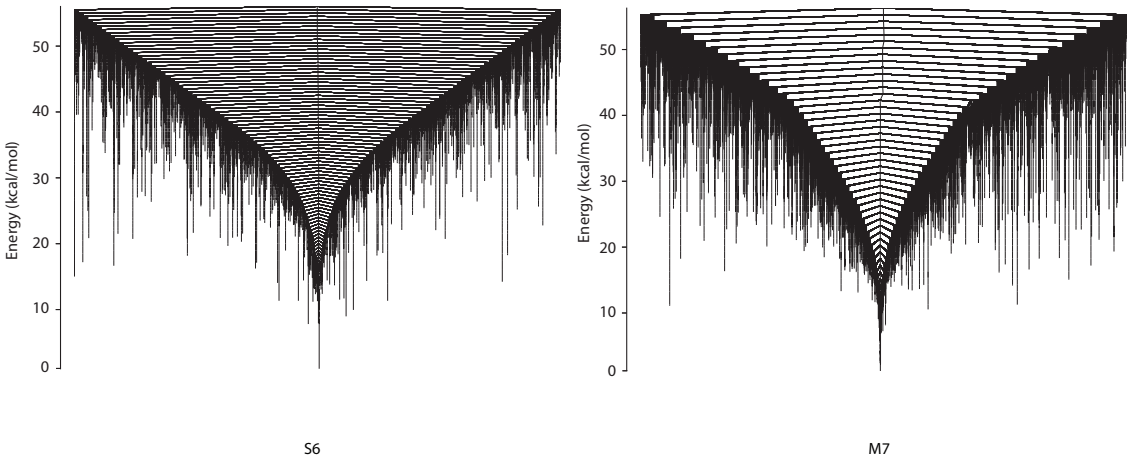

Figure 4: Disconnectivity graphs with the native Cα model (4.5 Å cut-off map) for proteins S6 and M7. No significant sub-funnels were found within our criterion of 50 nodes at $\delta = 0.5\epsilon$, indicating a highly funnelled structure-seeking landscape.

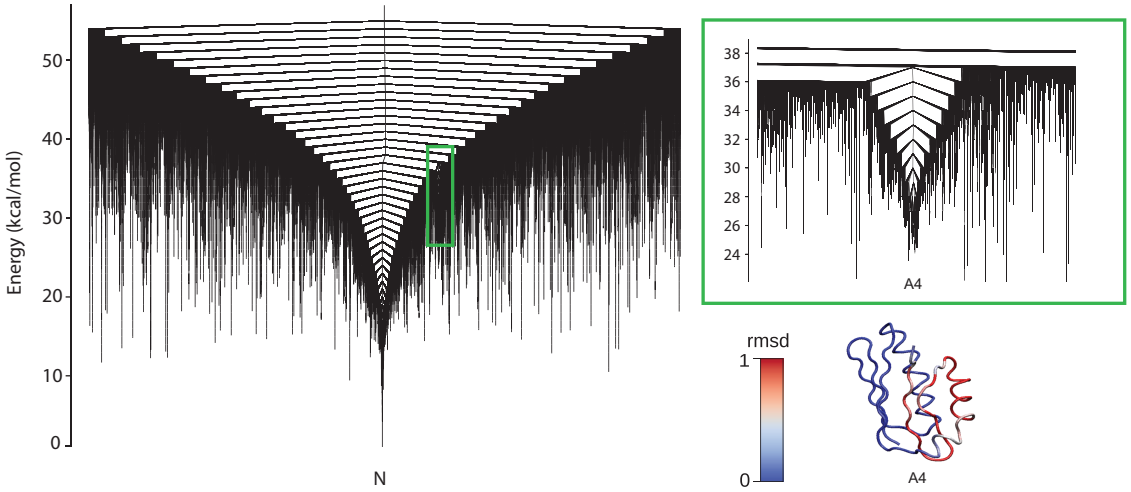

Figure 5: (a) Disconnectivity graph representation of the ensemble derived with the Cα-SBM with non-native hydrophobic interactions included. Thorough sampling up to 55 kcal/mol above the native state was performed with 403,111 minima 653,226 transition states. Conformer A4 corresponding to the bottom of one sub-funnel containing more than 50 nodes is depicted in the green box. Its structure is shown in the bottom right corner. The colouring is identical to Figure 3 .

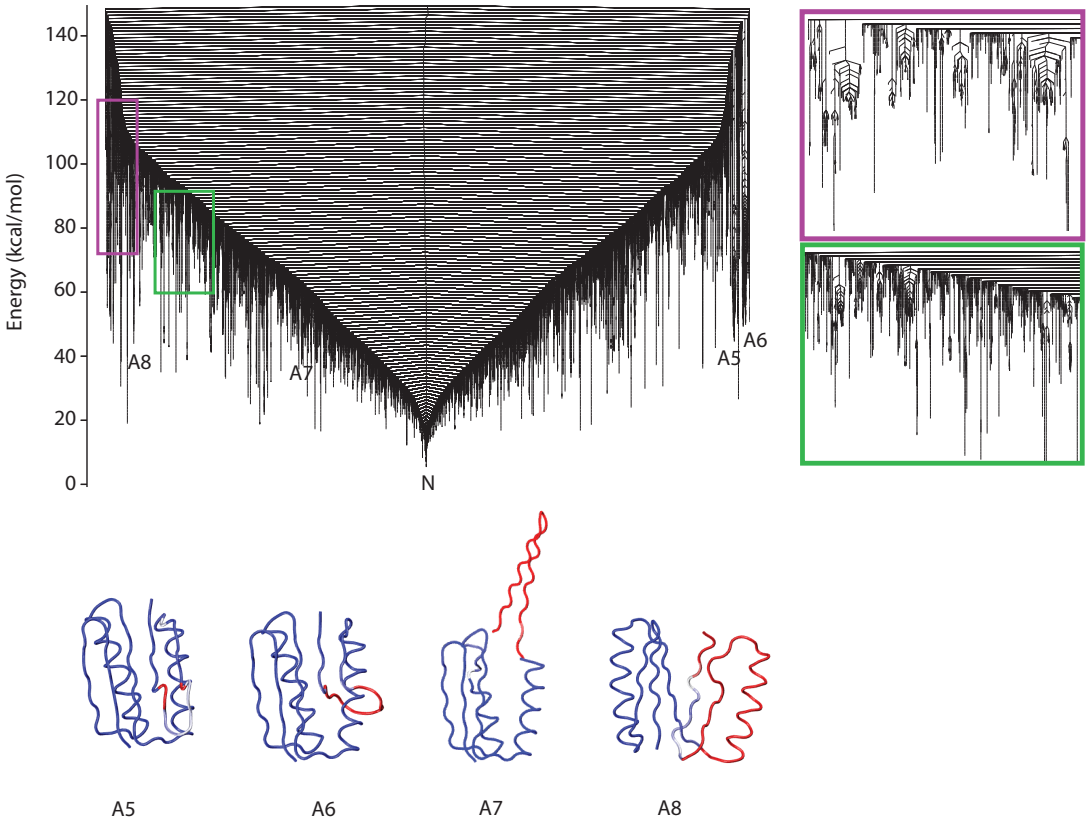

Figure 6: (a) Disconnectivity graph for the Cα-Cβ model. Two magnified sections of the graph emphasise the frustration. Many such sub-funnels with more than 50 nodes in a 0.5 ϵ energy range are found. The four conformers at the bottom of four biggest sub-funnels are labelled A5, A6, A7 and A8. The colouring is identical to Figure 3. 497,140 transition states connected to 225,893 local minima are contained in the graph.

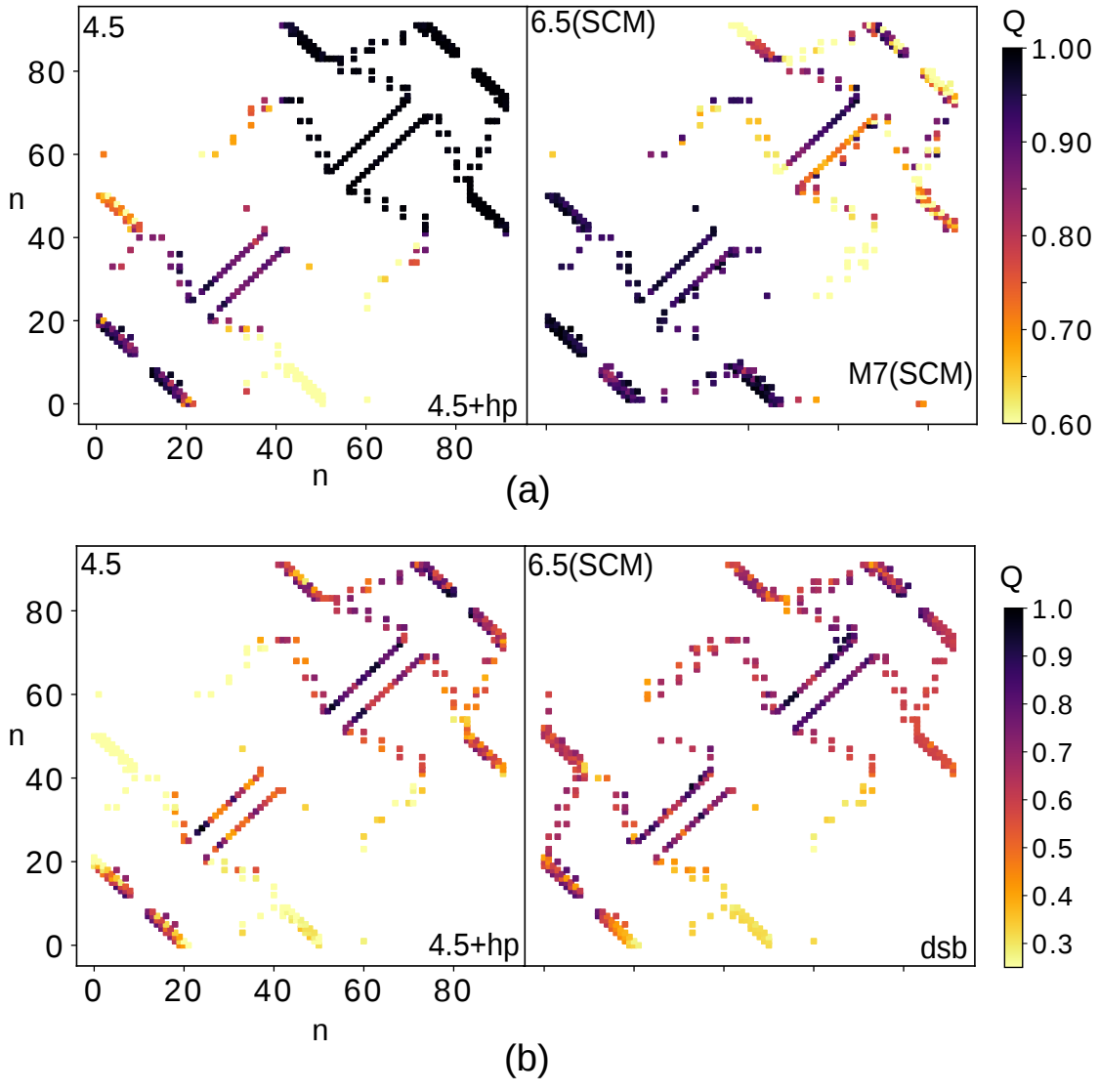

Figure 7: Probability contact maps for ensembles obtained from discrete path sampling[9,10] are depicted in (a). Four separate probability contact maps are plotted for potentials from the Cα-SBM with 4.5 Å cut-off (labelled 4.5), Cα-SBM with non-native hydrophobic interactions (labelled 4.5+hp), shadow contact map algorithm (SMOG server, labelled SCM) and for the M7 protein (SMOG server, labelled M7(SCM)). Similarly, in (b), four separate maps are plotted for ensembles derived from MD trajectories at around T_f for the Cα-SBM with a 4.5 Å cut-off (labelled 4.5), Cα-SBM with non-native hydrophobic interactions (labelled 4.5+hp),

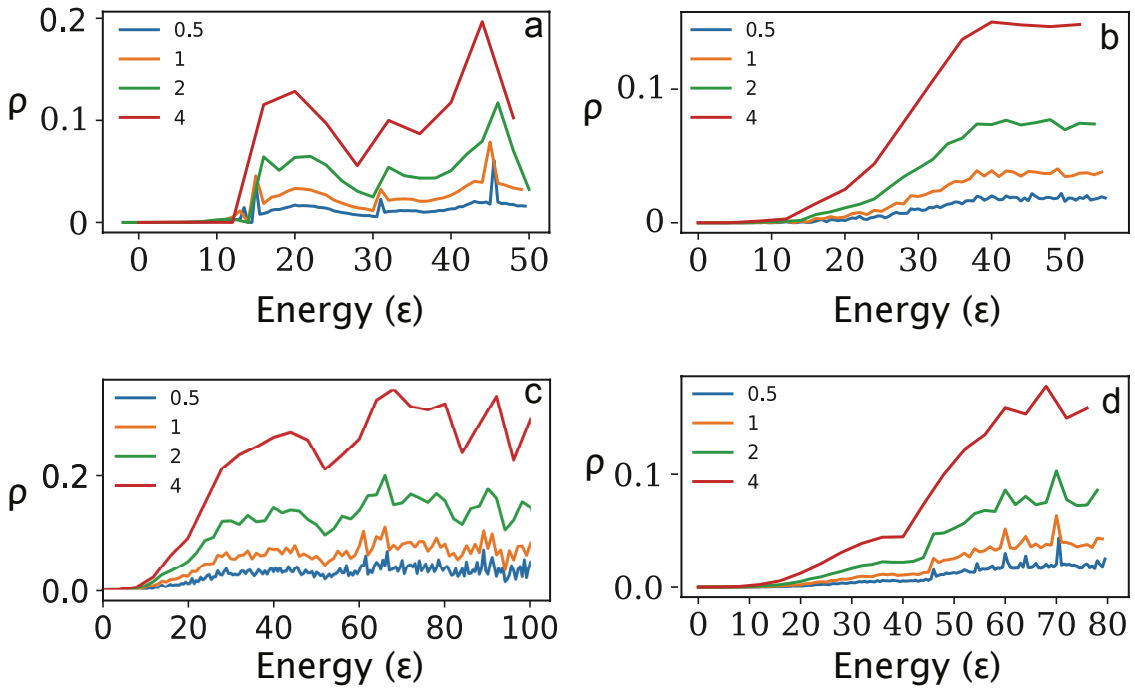

Figure 8: Frustration density ρ as a function of the energy interval, for disconnectivity graphs constructed with δ=0.5ϵ (blue), 1ϵ (orange), 2ϵ (green) and 4ϵ (red) for (a) Top7 - Cα model 4.5 Å cut-off based map, (b) S6 - Cα model, SCM-based map, (c) Top7 - Cα-Cβ model and (d) M7- Cα model, 6.5 Å SCM based map. The horizontal axis is scaled to δ=1 ϵ (kcal/mol). For the designed protein Top7, a funnelled structure-based model representation shows distinct fluctuation in the frustration density parameter. These fluctuations are absent for the naturally evolved ribosomal protein, S6, and the cooperative folder to Top7 topology, M7.

Acknowledgement

SN acknowledges support from National Centre for Biological Sciences (NCBS) through the NCBS/inSTEM-Cambridge (NiC) fellowship. DJW gratefully acknowledges support from the U.K. Engineering and Physical Sciences Research Council (EPSRC) for funding under grant EP/N035003/1. SN and SG acknowledge support by core funding from the Tata Institute of Fundamental Research (TIFR).

References

(1) Wales, D. J. Decoding the energy landscape: extracting structure, dynamics and thermodynamics. *Philos. Trans. Royal Soc. A* **2012**, *370*, 2877–2899.

(2) Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem* **1997**, *48*, 545–600.

(3) Onuchic, J. N.; Wolynes, P. G. Theory of protein folding. *Curr. Opin. Struct. Biol.* **2004**, *14*, 70–75.

(4) Levinthal, C. How to Fold Graciously. Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois. 1969; pp 22–24.

(5) Bryngelson, J. D.; Wolynes, P. G. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci.* **1987**, *84*, 7524–7528.

(6) Clementi, C. Coarse-grained models of protein folding: toy models or predictive tools? *Curr. Opin. Struct. Biol.* **2008**, *18*, 10–15.

(7) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins: Struct. Funct. Bioinf.* **1995**, *21*, 167–195.

(8) Clementi, C.; Nymeyer, H.; Onuchic, J. N. Topological and energetic factors: what determines the structural details of the transition state ensemble and en-route intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* **2000**, *298*, 937–953.

(9) Wales, D. J. Discrete path sampling. *Mol. Phys.* **2002**, *100*, 3285–3306.

(10) Wales, D. J. Some further applications of discrete path sampling to cluster isomerization. *Mol. Phys.* **2004**, *102*, 891–908.

(11) Becker, O. M.; Karplus, M. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *J. Chem. Phys.* **1997**, *106*, 1495–1517.

(12) Wales, D. J.; Miller, M. A.; Walsh, T. R. Archetypal energy landscapes. *Nature* **1998**, *394*, 758–760.

(13) Neelamraju, S.; Oakley, M. T.; Johnston, R. L. Chiral effects on helicity studied via the energy landscape of short (d, l)-alanine peptides. *J. Chem. Phys.* **2015**, *143*, 10B618_1.

(14) Wales, D. *Energy landscapes: Applications to clusters, biomolecules and glasses*; Cambridge University Press, 2003.

(15) Somani, S.; Wales, D. J. Energy landscapes and global thermodynamics for alanine peptides. *J. Chem. Phys.* **2013**, *139*, 121909.

(16) Carr, J. M.; Trygubenko, S. A.; Wales, D. J. Finding pathways between distant local minima. *J. Chem. Phys.* **2005**, *122*, 234903.

(17) Chakraborty, D.; Wales, D. J. Energy Landscape and Pathways for Transitions Between Watson-Crick and Hoogsteen Base Pairing in DNA. *J. Phys. Chem. Lett.* **2017**,

(18) Röder, K.; Wales, D. J. Evolved Minimal Frustration in Multifunctional Biomolecules. *J. Phys. Chem. B* **2018**,

(19) Chakraborty, D.; Collepardo-Guevara, R.; Wales, D. J. Energy Landscapes, Folding Mechanisms, and Kinetics of RNA Tetraloop Hairpins. *J. Am. Chem. Soc.* **2014**, *136*, 18052–18061, PMID: 25453221.

(20) Chebaro, Y.; Ballard, A. J.; Chakraborty, D.; Wales, D. J. Intrinsically disordered energy landscapes. *Sci. Rep.* **2015**, *5*, 10386.

(21) Röder, K.; Wales, D. J. Transforming the Energy Landscape of a Coiled-Coil Peptide via Point Mutations. *J. Chem. Theory Comput* **2017**, *13*, 1468–1477, PMID: 28177620.

(22) Oakley, M. T.; Wales, D. J.; Johnston, R. L. Energy landscape and global optimization for a frustrated model protein. *J. Phys. Chem. B* **2011**, *115*, 11525–11529.

(23) Levy, Y.; Wolynes, P. G.; Onuchic, J. N. Protein topology determines binding mechanism. *Proc. Natl. Acad. Sci.* **2004**, *101*, 511–516.

(24) Zhang, Z.; Chan, H. S. Native topology of the designed protein Top7 is not conducive to cooperative folding. *Biophys. J.* **2009**, *96*, L25–L27.

(25) Gosavi, S.; Chavez, L. L.; Jennings, P. A.; Onuchic, J. N. Topological frustration and the folding of interleukin-1β. *J. Mol. Biol.* **2006**, *357*, 986–996.

(26) Betancourt, M. R.; Onuchic, J. N. Kinetics of proteinlike models: the energy landscape factors that determine folding. *J. Chem. Phys.* **1995**, *103*, 773–787.

(27) Dantas, G.; Watters, A. L.; Lunde, B. M.; Eletr, Z. M.; Isern, N. G.; Roseman, T.; Lipfert, J.; Doniach, S.; Tompa, M.; Kuhlman, B.; Stoddard, B. L.; Varani, G.; Baker, D. Mis-translation of a Computationally Designed Protein Yields an Exceptionally Stable Homodimer: Implications for Protein Engineering and Evolution. *J. Mol. Biol.* **2006**, *362*, 1004 – 1024.

(28) Scalley-Kim, M.; Baker, D. Characterization of the Folding Energy Landscapes of Computer Generated Proteins Suggests High Folding Free Energy Barriers and Cooperativity may be Consequences of Natural Selection. *J. Mol. Biol.* **2004**, *338*, 573 – 583.

(29) Yadahalli, S.; Hemanth Giri Rao, V.; Gosavi, S. Modeling Non-Native Interactions in Designed Proteins. *Isr. J. Chem.* **2014**, *54*, 1230–1240.

(30) Zhang, Z.; Chan, H. S. Competition between native topology and nonnative interactions in simple and complex folding kinetics of natural and designed proteins. *Proc. Natl. Acad. Sci.* **2010**, *107*, 2920–2925.

(31) Funneling and frustration in the energy landscapes of some designed and simplified proteins. *J. Chem. Phys.* **2013**, *139*, 121908.

(32) Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D. Design of a novel globular protein fold with atomic-level accuracy. *Science* **2003**, *302*, 1364–1368.

(33) Watters, A. L.; Deka, P.; Corrent, C.; Callender, D.; Varani, G.; Sosnick, T.; Baker, D. The highly cooperative folding of small naturally occurring proteins is likely the result of natural selection. *Cell* **2007**, *128*, 613–624.

(34) Lindahl, M.; Svensson, L.; Liljas, A.; Sedelnikova, S.; Eliseikina, I.; Fomenkova, N.; Nevskaya, N.; Nikonov, S.; Garber, M.; Muranova, T. Crystal structure of the ribosomal protein S6 from Thermus thermophilus. *The EMBO Journal* **1994**, *13*, 1249–1254.

(35) Yadahalli, S.; Gosavi, S. Designing cooperativity into the designed protein Top7. *Proteins: Struct. Funct. Bioinf.* **2014**, *82*, 364–374.

(36) Dallüge, R.; Oschmann, J.; Birkenmeier, O.; Lücke, C.; Lilie, H.; Rudolph, R.; Lange, C. A tetrapeptide fragment-based design method results in highly stable artificial proteins. *Proteins: Struct. Funct. Bioinf.* **2007**, *68*, 839–849.

(37) Stordeur, C.; Dallüge, R.; Birkenmeier, O.; Wienk, H.; Rudolph, R.; Lange, C.; Lücke, C. The NMR solution structure of the artificial protein M7 matches the computationally designed model. *Proteins: Struct. Funct. Bioinf.* **2008**, *72*, 1104–1107.

(38) Levy, Y.; Becker, O. M. Effect of Conformational Constraints on the Topography of Complex Potential Energy Surfaces. *Phys. Rev. Lett.* **1998**, *81*, 1126–1129.

(39) Noel, J. K.; Levi, M.; Raghunathan, M.; Lammert, H.; Hayes, R. L.; Onuchic, J. N.; Whitford, P. C. SMOG 2: A Versatile Software Package for Generating Structure-Based Models. *PLOS Computational Biology* **2016**, *12*, 1–14.

(40) Sutto, L.; Lätzer, J.; Hegler, J. A.; Ferreiro, D. U.; Wolynes, P. G. Consequences of localized frustration for the folding mechanism of the IM7 protein. *Proc. Natl. Acad. Sci.* **2007**, *104*, 19825–19830.

(41) Liu, Z.; Chan, H. S. Solvation and desolvation effects in protein folding: native flexibility, kinetic cooperativity and enthalpic barriers under isostability conditions. *Phys. Biol.* **2005**, *2*, S75.

(42) Rank, J. A.; Baker, D. A desolvation barrier to hydrophobic cluster formation may contribute to the ratelimiting step in protein folding. *Prot. Sci.* **2001**, *6*, 347–354.

(43) Cheung, M. S.; Finke, J. M.; Callahan, B.; Onuchic, J. N. Exploring the Interplay between Topology and Secondary Structural Formation in the Protein Folding Problem. *J. Phys. Chem. B.* **2003**, *107*, 11193–11200.

(44) Nocedal, J. Updating quasi-Newton matrices with limited storage. *Math. Comput.* **1980**, *35*, 773–782.

(45) Wales, D. OPTIM: A program for optimising geometries and calculating pathways. *See http://www-wales. ch. cam. ac. uk/OPTIM* **1999**, Last accessed: 28th Feb, 2018.

(46) Wales, D. PATHSAMPLE: A program for generating connected stationary point databases and extracting global kinetics. *See http://www-wales. ch. cam. ac. uk/PATHSAMPLE* **1999**, Last accessed: 28th Feb, 2018.

(47) Carr, J. M.; Trygubenko, S. A.; Wales, D. J. Finding pathways between distant local minima. *J. Chem. Phys.* **2005**, *122*, 234903.

(48) Trygubenko, S. A.; Wales, D. J. A Doubly Nudged Elastic Band Method for Finding Transition States. *J. Chem. Phys.* **2004**, *120*, 2082–2094.

(49) Sheppard, D.; Terrell, R.; Henkelman, G. Optimization methods for finding minimum energy paths. *J. Chem. Phys.* **2008**, *128*, 134106.

(50) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.* **2000**, *113*, 9901–9904.

(51) Henkelman, G.; Jónsson, H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.* **2000**, *113*, 9978–9985.

(52) Wales, D. J.; Carr, J. M. Quasi-Continuous Interpolation Scheme for Pathways between Distant Configurations. *J. Chem. Theory Comput* **2012**, *8*, 5020–5034.

(53) Munro, L. J.; Wales, D. J. Defect migration in crystalline silicon. *Phys. Rev. B* **1999**, *59*, 3969.

(54) Krivov, S. V.; Karplus, M. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Natl. Acad. Sci.* **2004**, *101*, 14766–14770.

(55) Smeeton, L. C.; Farrell, J. D.; Oakley, M. T.; Wales, D. J.; Johnston, R. L. Structures and energy landscapes of hydrated sulfate clusters. *J. Chem. Theory Comput* **2015**, *11*, 2377–2384.

(56) Derrida, B. Random-energy model: An exactly solvable model of disordered systems. *Phys. Rev. B* **1981**, *24*, 2613–2626.

(57) Parra, R. G.; Schafer, N. P.; Radusky, L. G.; Tsai, M.-Y.; Guzovsky, A. B.; Wolynes, P. G.; Ferreiro, D. U. Protein Frustratometer 2: a tool to localize energetic frustration in protein molecules, now with electrostatics. *Nucleic Acids Res.* **2016**, gkw304.

(58) Wolynes, P. G. Evolution, energy landscapes and the paradoxes of protein folding. *Biochimie* **2015**, *119*, 218 – 230.

(59) Norcross, T. S.; Yeates, T. O. A framework for describing topological frustration in models of protein folding. *J. Mol. Biol.* **2006**, *362*, 605–621.

(60) de Souza, V. K.; Stevenson, J. D.; Niblett, S. P.; Farrell, J. D.; Wales, D. J. Defining and quantifying frustration in the energy landscape: Applications to atomic and molecular clusters, biomolecules, jammed and glassy systems. *J. Chem. Phys.* **2017**, *146*, 124103.

(61) Mohanty, S.; Meinke, J. H.; Zimmermann, O.; Hansmann, U. H. Simulation of Top7-CFr: A transient helix extension guides folding. *Proc. Nat. Acad. Sci.* **2008**, *105*, 8004–8007.

(62) Ejtehadi, M. R.; Avall, S. P.; Plotkin, S. S. Three-body interactions improve the prediction of rate and mechanism in protein folding models. *Proc. Natl. Acad. Sci.* **2004**, *101*, 15088–15093.

(63) Craig, P. O.; Laetzer, J.; Weinkam, P.; Hoffman, R. M. B.; Ferreiro, D. U.; Komives, E. A.; Wolynes, P. G. Prediction of Native-State Hydrogen Exchange from Perfectly Funneled Energy Landscapes. *J. Am. Chem. Soc.* **2011**, *133*, 17463–17472, PMID: 21913704.

(64) Chen, T.; Chan, H. S. Native contact density and nonnative hydrophobic effects in the folding of bacterial immunity proteins. *PLoS Comput Biol* **2015**, *11*, e1004260.

(65) Ferreiro, D. U.; Hegler, J. A.; Komives, E. A.; Wolynes, P. G. Localizing frustration in native proteins and protein assemblies. *Proc. Natl. Acad. Sci.* **2007**, *104*, 19819–19824.

(66) Chen, M.; Lin, X.; Lu, W.; Onuchic, J. N.; Wolynes, P. G. Protein Folding and Structure Prediction from the Ground Up II: AAWSEM for $\alpha\beta$Proteins. *J. Phys. Chem. B* **2017**, *121*, 3473–3482.

(67) Sharma, D.; Perisic, O.; Peng, Q.; Cao, Y.; Lam, C.; Lu, H.; Li, H. Single-molecule force spectroscopy reveals a mechanically stable protein fold and the rational tuning of its mechanical stability. *Proc. Natl. Acad. Sci.* **2007**, *104*, 9278–9283.

Supporting Information Available

Potentials

The coarse-grained description of a flexible protein chain is as follows: Each residue is approximated as a spherical bead centered on the α -Carbon. N is the number of Cα beads in a given protein, r is the distance between two adjacent beads, θ the angle defined by three consecutive beads and ϕ the dihedral angle defined over four consecutive beads. U_{bond}, U_{angle} and $U_{dihedral}$ (summed up as U_1) are the energy contributions from these bonds, angles and dihedrals respectively. r_O, θ_O and ϕ_O are the corresponding bond lengths, angles and dihedrals in the native structure derived from the Protein Data Bank (PDB). Interactions between residues that are in contact in the native structure are given by U_{native} where r_{ij} is the distance between two Cα beads, i and j separated by at least three consecutive residues. If the Cα atoms are not in contact, then the excluded-volume repulsion for non-native contacts

is given by U_{rep}. Equation 1 summarises the definitions.

$$\begin{aligned} U_{bond} &= \sum_{i=1,N-1} \frac{K_b}{2}\big(r_i - r_O\big)^2 \\ U_{angle} &= \sum_{i=1,N-2} K_\theta\big(\theta_i - \theta_O\big)^2 \\ U_{dihedral} &= \sum_{i=1,N-3} \frac{K_\phi^{(n)}}{2}\big(1 - cos\big(n(\phi_i - \phi_O\big)\big) \\ U_1 &= U_{bond} + U_{angle} + U_{dihedral} \\ U_{native} &= \sum_{i<j-3}^{Native} \epsilon_{ij}\Big[5\Big(\frac{\sigma_{ij}}{r_{ij}}\Big)^{12} - 6\Big(\frac{\sigma_{ij}}{r_{ij}}\Big)^{10}\Big] \\ U_{rep} &= \sum_{i<j-3}^{Non-native} \epsilon_{ij}\Big(\frac{C}{r_{ij}}\Big)^{12} \end{aligned} \tag{1}$$

The potential for the Cα-SBM is then defined in equation 2.

$$U_{C\alpha} = U_1 + U_{native} + U_{rep} \tag{2}$$

where $K_b = 100\epsilon$, $K_\theta = 20\epsilon$, $K_\phi^{(1)} = \epsilon$, $K_\phi^{(3)} = 0.5\epsilon$ C=4 Å ,ϵ_{ij}=1 kcal/mol. More details in Ref.[39]

When hydrophobic contacts are included, denoted by "hp" in the manuscript, the potential employed is given by equation 3.

$$U_{hp} = U_{C\alpha} + \sum_{i<j-3}^{HP-Non-native} \epsilon_{ij}\Big[5\Big(\frac{\sigma_{hp}}{r_{ij}}\Big)^{12} - 6\Big(\frac{\sigma_{hp}}{r_{ij}}\Big)^{10}\Big] \tag{3}$$

Each hydrophobic contact is given a default equilibrium distance of $\sigma_{hp} = 5.5$ Å and an interaction strength of 1ϵ. "HP-Non-native" refers to summation of pairs of hydrophobic residues that are not already in contact in the native state.

The desolvation barrier potential that allows a coarse-grained approximation of a water

molecule embedded within the protein denoted by "dsb" in the manuscript is defined in equation 4:

$$U_{dsb} = U_1 + U_2 + U_{rep} \tag{4}$$

where U_2 is the attractive interaction among the Cα beads in contact in the native state and is defined in equation 5 as:

$$U_2 = \begin{cases} \epsilon Z(r)\big[Z(r-2)\big], & \mathrm{r}<\mathrm{r}_{cm}. \\ CY(r^n)\big[Y(r)^n/2 - (r_{db}-r_{cm})^{2n}/2n + \epsilon_{db}\big], & \mathrm{r}_{cm} \le r < r_{db}. \\ -B\big[Y(r)-h_1\big]/\big[Y(r)^m + h_2\big], & r > r_{db} \end{cases} \tag{5}$$

where:

$$\begin{aligned} Z(r) &= (r_{cm}/r)^k \\ Y(r) &= (r - r_{db})^2 \\ C &= 4n(\epsilon + \epsilon_{db})/(r_{db} - r_{cm})^{4n} \\ B &= m\epsilon_{ssm}(r_{ssm} - r_{db})^{2(m-1)} \\ h_1 &= (1-1/m)(r_{ssm} - r_{db})^2/(\epsilon_{ssm}/\epsilon_{db} + 1) \\ h_2 &= (m-1)(r_{ssm} - rdb)^{2m}/(1 + \epsilon_{db}/\epsilon_{ssm}) \\ k_{db} &= 6, m_{db} = 3, n_{db} = 2 \end{aligned} \tag{6}$$

Both U_2 and its first derivative with respect to r_{ij} are continuous functions. More details in Reference.[41,42]

The two-bead potential was introduced by Cheung et al.[43] We have employed the most basic form that does not account for chirality and hydrogen bonding. We use the potential only to show a clear increase in frustration on introducing coarse-grained side-chain interactions.

$$U_{tb} = U_{bond} + U_{angle} + \sum_{i=1,N-3}^{C_\alpha - C_\alpha - C_\alpha - C_\alpha} \frac{K_\phi}{2}\big(1 - cos\big(2\phi_i - \frac{\pi}{2}\big)\big) + U_{native} + \sum_{i<j-3}^{Non-native} \epsilon_{ij}\Big(\frac{f\sigma_{ij}}{r_{ij}}\Big)^{12} \tag{7}$$

where i and j are indices of beads. $\sigma_{ij} = r_i + r_j$, where r_i and r_j are the radii of interacting beads. For non-native Cα-Cβ pairs,$|i-j| \geq 2$; for non-native Cα-Cα pairs,$|i-j| \geq 4$ and for non-native Cβ-Cβ pairs; $|i-j| \geq 2$ for non-native and pairs."f" is the scaling factor set at 0.7. More details of the model can be found in reference.[43]

Table 1: List of SBM flavours and parameters for each protein studied. Total number of minima (n_{min}) and transition states (n_{trans}) found for each type of SBM after PATHSAMPLE runs. K_b,K_θ and K_ϕ are the harmonic constants for bonds, angles and dihedrals respectively as described in eq1.

Protein	Type of SBM	r_c	$n_{contacts}$	n_{min}, n_{ts}	K_b,K_θ,K_ϕ
Top7	Cα(SCM)	6.5	261	526166,611648	100,20,1
Top7	Cα(cut-off)	4.5	201	506037,555854	100,20,1
Top7	Cα(hp)	4.5	201+659	403111,653226	100,20,1
Top7	Cα(dsb)	4.5	201	n/a	100,20,1
Top7	Cα-Cβ	4.5	402	225893,497140	100,20,1
S6	Cα(SCM)	6.5	282	50803,68066	100,20,1
S6	Cα(cut-off)	4.5	211	79628,91290	100,20,1
S6	Cα(hp)	4.5	211+853	45470,47148	100,20,1
S6	Cα(dsb)	4.5	211	n/a	100,20,1
S6	Cα-Cβ	4.5	402	40458,49737	100,20,1
M7	Cα(SCM)	6.5	229	320609,312942	100,20,1
M7	Cα(cut-off)	4.5	253	475011,418546	100,20,1

Gō-kit commands

In order to determine the residues in contact in the native state, two separate methods were employed. The Shadow Contact Map (SCM) algorithm[3] and a cut-off based contact map generated from the Gō-kit code maintained by us [4]. Parameters for each potential are summarised in Table 1.

[3]https://smog-server.org

[4]https://bitbucket.org/nsridhar/pyesbm

Contact maps

Top7-4.5 Å cut-off (201 contacts)

1 20 ; 1 21 ; 1 22 ; 1 51 ; 2 19 ; 2 20 ; 2 21 ; 2 50 ; 2 51 ; 2 61 ; 3 18 ; 3 19 ; 3 20 ; 3 49 ; 3 50 ; 3 51 ; 4 17 ; 4 18 ; 4 19 ; 4 34 ; 4 48 ; 4 49 ; 4 50 ; 5 16 ; 5 17 ; 5 18 ; 5 47 ; 5 48 ; 5 49 ; 6 15 ; 6 16 ; 6 17 ; 6 19 ; 6 34 ; 6 47 ; 6 48 ; 7 14 ; 7 15 ; 7 16 ; 7 45 ; 7 46 ; 7 47 ; 8 13 ; 8 14 ; 8 15 ; 8 17 ; 8 38 ; 8 43 ; 8 45 ; 8 46 ; 9 13 ; 9 14 ; 9 43 ; 9 44 ; 9 45 ; 10 41 ; 10 43 ; 10 44 ; 13 41 ; 15 41 ; 17 34 ; 17 37 ; 19 30 ; 19 33 ; 19 34 ; 19 37 ; 21 26 ; 21 27 ; 21 30 ; 22 26 ; 24 28 ; 24 61 ; 25 29 ; 26 30 ; 27 31 ; 27 61 ; 28 32 ; 29 33 ; 30 34 ; 31 35 ; 31 64 ; 31 65 ; 32 36 ; 33 37 ; 34 38 ; 34 48 ; 35 39 ; 35 68 ; 35 71 ; 35 72 ; 36 40 ; 37 41 ; 38 42 ; 38 43 ; 38 72 ; 38 74 ; 39 72 ; 42 74 ; 42 92 ; 43 74 ; 43 92 ; 44 74 ; 44 90 ; 44 91 ; 44 92 ; 45 90 ; 45 91 ; 46 69 ; 46 74 ; 46 88 ; 46 89 ; 46 90 ; 47 87 ; 47 88 ; 47 89 ; 48 65 ; 48 69 ; 48 86 ; 48 87 ; 48 88 ; 49 85 ; 49 86 ; 49 87 ; 50 61 ; 50 65 ; 50 84 ; 50 85 ; 50 86 ; 51 84 ; 51 85 ; 52 57 ; 52 58 ; 52 61 ; 52 84 ; 53 57 ; 53 84 ; 54 58 ; 54 84 ; 55 59 ; 55 81 ; 55 83 ; 55 84 ; 56 60 ; 57 61 ; 58 62 ; 58 81 ; 58 84 ; 58 86 ; 59 63 ; 59 81 ; 60 64 ; 61 65 ; 61 86 ; 62 66 ; 62 79 ; 62 81 ; 62 86 ; 62 88 ; 63 67 ; 64 68 ; 65 69 ; 65 88 ; 66 70 ; 66 77 ; 66 79 ; 67 71 ; 68 72 ; 69 73 ; 69 74 ; 69 77 ; 69 88 ; 69 89 ; 69 90 ; 70 74 ; 70 75 ; 70 77 ; 72 92 ; 73 91 ; 73 92 ; 74 90 ; 74 91 ; 74 92 ; 75 91 ; 75 92 ; 76 89 ; 76 90 ; 76 91 ; 77 88 ; 77 89 ; 77 90 ; 78 87 ; 78 88 ; 78 89 ; 79 86 ; 79 87 ; 79 88 ; 80 85 ; 80 86 ; 80 87 ; 81 85 ; 81 86

Top7-6.5 Å Shadow Contact-Map (261 contacts)

1 53 ; 1 51 ; 1 52 ; 1 20 ; 1 21 ; 1 22 ; 2 20 ; 2 21 ; 2 22 ; 2 51 ; 2 52 ; 2 19 ; 2 53 ; 2 57 ; 2 61 ; 2 27 ; 2 50 ; 2 23 ; 2 24 ; 3 50 ; 3 51 ; 3 19 ; 3 20 ; 3 49 ; 3 18 ; 4 18 ; 4 19 ; 4 21 ; 4 48 ; 4 49 ; 4 50 ; 4 17 ; 4 30 ; 4 34 ; 4 27 ; 5 48 ; 5 49 ; 5 50 ; 5 17 ; 5 18 ; 5 47 ; 5 16 ; 6 16 ; 6 17 ; 6 19 ; 6 47 ; 6 48 ; 6 15 ; 6 34 ; 6 38 ; 6 46 ; 6 30 ; 7 46 ; 7 47 ; 7 14 ; 7 15 ; 7 16 ; 7 45 ; 8 14 ; 8 15 ; 8 17 ; 8 45 ; 8 46 ; 8 13 ; 8 47 ; 8 38 ; 8 41 ; 8 43 ; 8 37 ; 9 43 ; 9 45 ; 9 46 ; 9 13 ; 9 14 ; 9 41 ; 9 44 ; 10 44 ; 10 41 ; 10 43 ; 10 42 ; 13 41 ; 15 41 ; 17 34 ; 17 37 ; 19 30 ; 19 34 ; 19 33 ; 19 37 ; 20 26 ; 21 26 ; 21 27 ; 21 30 ; 22 26 ; 23 27 ; 24 61 ; 24 28 ; 24 64 ; 24 60 ; 25 29 ; 26 30 ; 27 31 ; 27 50 ; 27 61 ; 27 64 ; 28 32 ; 28 33 ; 29 33 ; 30 34 ; 31 35 ; 31 68 ; 31 64 ; 31 65 ; 31 50 ; 32 36 ; 33 37 ; 34 38 ; 34 39 ; 34 72 ; 34 48 ; 35 68 ; 35 72 ; 35 39 ; 35 67 ; 35 71 ; 36 40 ; 36 41 ; 37 41 ; 37 42 ; 38 74 ; 38 42 ; 38 43 ; 38 46 ; 38 72 ; 38 48 ; 39 72 ; 39 71 ; 42 74 ; 42 92 ; 43 74 ; 43 92 ; 44 92 ; 44 74 ; 44 90 ; 44 91 ; 45 89 ; 45 90 ; 45 91 ; 46 74 ; 46 89 ; 46 90 ; 46 69 ; 46 88 ; 47 88 ; 47 89 ; 47 87 ; 48 69 ; 48 87 ; 48 88 ; 48 65 ; 48 86 ; 48 68 ; 49 86 ; 49 87 ; 49 85 ; 50 65 ; 50 85 ; 50 86 ; 50 58 ; 50 61 ; 50 84 ; 51 84 ; 51 85 ; 51 61 ; 52 58 ; 52 84 ; 52 85 ; 52 86 ; 52 57 ; 52 61 ; 53 57 ; 53 84 ; 54 84 ; 54 58 ; 55 84 ; 55 81 ; 55 83 ; 55 59 ; 56 60 ; 57 61 ; 58 86 ; 58 62 ; 58 81 ; 58 83 ; 58 84 ; 59 63 ; 59 81 ; 60 64 ; 61 65 ; 61 88 ; 61 86 ; 62 86 ; 62 88 ; 62 66 ; 62 79 ; 62 81 ; 63 67 ; 64 68 ; 65 88 ; 65 69 ; 65 70 ; 65 77 ; 65 86 ; 66 77 ; 66 70 ; 66 74 ; 66 79 ; 66 88 ; 67 71 ; 67 72 ; 68 72 ; 68 74 ; 69 74 ; 69 73 ; 69 75 ; 69 77 ; 69 90 ; 69 88 ; 69 89 ; 70 77 ; 70 74 ; 70 75 ; 72 92 ; 73 91 ; 73 92 ; 74 91 ; 74 92 ; 74 90 ; 75 91 ; 75 92 ; 76 90 ; 76 91 ; 76 92 ; 76 89 ; 77 89 ; 77 90 ; 77 91 ; 77 88 ; 78 88 ; 78 89 ; 78 87 ; 78 91 ; 79 86 ; 79 87 ; 79 88 ; 80 86 ; 80 87 ; 80 88 ; 80 85 ; 81 85 ; 81 86 ; 82 87

Top7-Cα-Cβ;4.5 Å cut-off (402 contacts)

1 99 ; 2 43 ; 2 39 ; 3 40 ; 3 38 ; 3 41 ; 3 39 ; 3 36 ; 3 99 ; 4 8 ; 4 41 ; 4 98 ; 4 120 ; 5 99 ; 5 100 ; 5 98 ; 5 36 ; 5 95 ; 5 97 ; 6 39 ; 6 35 ; 6 100 ; 7 32 ; 7 98 ; 7 37 ; 7 35 ; 7 95 ; 7 36 ; 7 34 ; 8 37 ; 8 12 ; 8 67 ; 8 94 ; 8 98 ; 9 96 ; 9 94 ; 9 93 ; 9 91 ; 9 95 ; 9 32 ; 10 96 ; 10 35 ; 10 31 ; 11 33 ; 11 31 ; 11 28 ; 11 30 ; 11 32 ; 11 91 ; 12 16 ; 12 94 ; 12 67 ; 12 37 ; 12 33 ; 13 92 ; 13 28 ; 13 87 ; 13 89 ; 13 91 ; 14 92 ; 14 27 ; 14 31 ; 15 24 ; 15 87 ; 15 28 ; 15 26 ; 15 29 ; 15 27 ; 16 29 ; 16 33 ; 16 75 ; 16 84 ; 16 90 ; 17 27 ; 17 83 ; 17 85 ; 17 84 ; 17 87 ; 17 24 ; 18 27 ; 19 24 ; 19 25 ; 20 81 ; 20 84 ; 21 25 ; 24 29 ; 25 29 ; 25 81 ; 29 33 ; 29 81 ; 30 35 ; 31 35 ; 32 37 ; 33 73 ; 33 67 ; 33 37 ; 37 67 ; 37 65 ; 37 59 ; 37 73 ; 40 51 ; 41 53 ; 41 59 ; 41 51 ; 42 51 ; 43 51 ; 44 49 ; 44 51 ; 44 50 ; 45 51 ; 45 49 ; 46 52 ; 46 54 ; 46 55 ; 46 53 ; 47 53 ; 47 120 ; 48 53 ; 48 57 ; 48 55 ; 48 56 ; 48 54 ; 50 58 ; 50 56 ; 50 57 ; 50 59 ; 51 57 ; 52 60 ; 52 58 ; 52 59 ; 52 61 ; 53 61 ; 53 120 ; 54 62 ; 54 60 ; 54 61 ; 54 63 ; 56 64 ; 56 62 ; 56 63 ; 56 65 ; 58 66 ; 58 64 ; 58 65 ; 58 67 ; 59 67 ; 60 68 ; 60 69 ; 60 67 ; 60 66 ; 61 128 ; 61 126 ; 62 71 ; 62 70 ; 62 68 ; 62 69 ; 63 71 ; 64 72 ; 64 70 ; 64 73 ; 64 71 ; 65 73 ; 66 72 ; 66 75 ; 66 73 ; 66 74 ; 67 73 ; 67 94 ; 68 77 ; 68 74 ; 68 76 ; 68 142 ; 68 75 ; 69 134 ; 69 140 ; 69 142 ; 70 76 ; 70 78 ; 70 77 ; 70 79 ; 71 77 ; 72 81 ; 72 78 ; 72 79 ; 72 80 ; 74 82 ; 74 80 ; 74 83 ; 74 145 ; 74 81 ; 75 145 ; 75 142 ; 76 82 ; 77 142 ; 81 84 ; 82 145 ; 83 145 ; 84 145 ; 84 90 ; 85 178 ; 85 176 ; 85 175 ; 85 145 ; 85 177 ; 86 177 ; 87 177 ; 87 175 ; 88 177 ; 89 136 ; 89 174 ; 89 175 ; 89 173 ; 89 171 ; 90 136 ; 90 145 ; 90 94 ; 91 171 ; 91 174 ; 91 173 ; 92 170 ; 92 174 ; 93 167 ; 93 171 ; 93 128 ; 93 136 ; 93 170 ; 93 172 ; 93 169 ; 94 98 ; 94 128 ; 94 136 ; 95 170 ; 95 167 ; 95 169 ; 96 166 ; 96 170 ; 97 166 ; 97 167 ; 97 163 ; 97 165 ; 97 168 ; 98 168 ; 98 128 ; 98 120 ; 99 166 ; 99 165 ; 99 163 ; 100 166 ; 101 163 ; 101 164 ; 101 112 ; 102 112 ; 102 120 ; 102 114 ; 103 112 ; 103 164 ; 104 112 ; 105 113 ; 105 111 ; 105 164 ; 105 114 ; 105 112 ; 106 110 ; 106 112 ; 107 162 ; 107 115 ; 107 113 ; 107 164 ; 107 116 ; 107 114 ; 108 159 ; 108 164 ; 108 116 ; 109 117 ; 109 116 ; 109 118 ; 109 115 ; 110 116 ; 111 119 ; 111 117 ; 111 118 ; 111 120 ; 112 120 ; 113 121 ; 113 119 ; 113 120 ; 113 159 ; 113 168 ; 113 122 ; 114 159 ; 114 168 ; 115 121 ; 115 123 ; 115 122 ; 115 124 ; 115 159 ; 116 159 ; 117 123 ; 117 124 ; 117 126 ; 117 125 ; 119 125 ; 119 127 ; 119 128 ; 119 126 ; 120 126 ; 120 168 ; 121 128 ; 121 130 ; 121 168 ; 121 172 ; 121 129 ; 121 127 ; 122 168 ; 122 155 ; 122 159 ; 123 129 ; 123 130 ; 123 132 ; 123 131 ; 125 133 ; 125 131 ; 125 134 ; 125 132 ; 126 134 ; 127 135 ; 127 133 ; 127 134 ; 127 136 ; 128 172 ; 129 137 ; 129 135 ; 129 151 ; 129 136 ; 129 138 ; 130 155 ; 131 138 ; 131 139 ; 131 137 ; 131 140 ; 133 141 ; 133 139 ; 133 142 ; 133 140 ; 134 142 ; 135 143 ; 135 141 ; 135 145 ; 135 142 ; 135 144 ; 136 151 ; 136 172 ; 136 145 ; 137 147 ; 137 151 ; 137 143 ; 137 144 ; 141 145 ; 142 145 ; 143 147 ; 143 176 ; 143 178 ; 144 151 ; 144 176 ; 144 178 ; 146 151 ; 146 178 ; 146 176 ; 148 173 ; 148 175 ; 148 176 ; 148 177 ; 149 177 ; 150 173 ; 150 175 ; 151 172 ; 151 155 ; 152 174 ; 152 173 ; 152 171 ; 152 169 ; 152 172 ; 153 174 ; 154 171 ; 154

169 ; 155 172 ; 155 168 ; 155 159 ; 156 170 ; 156 169 ; 156 167 ; 156 165 ; 156 168 ; 157 170 ; 158 167 ; 158 165 ; 159 168 ; 160 165 ; 160 166 ; 166 170 ; 168 172 ; 174 177 ;

S6-Cα;4.5 Åcut-off;211 contacts

1 34; 1 36; 1 66; 1 67; 1 68; 2 65; 2 66; 2 67; 2 69; 3 36; 3 38; 3 64; 3 65; 3 66; 3 92; 3 93; 3 96; 4 63; 4 64; 4 65; 4 67; 4 69; 4 72; 4 90; 4 91; 4 92; 4 93; 5 62; 5 63; 5 89; 5 90; 5 91; 5 93; 6 61; 6 62; 6 63; 6 65; 6 79; 6 88; 6 89; 6 90; 7 61; 7 62; 7 87; 7 88; 7 89; 7 91; 8 26; 8 59; 8 60; 8 61; 8 63; 8 79; 8 85; 8 87; 8 88; 9 48; 9 58; 9 59; 9 60; 9 84; 9 85; 9 86; 9 87; 10 19; 10 22; 10 58; 10 59; 10 61; 10 84; 10 85; 10 86; 11 59; 11 84; 11 86; 12 45; 12 55; 12 57; 12 58; 12 59; 12 86; 13 18; 13 55; 14 18; 14 19; 14 22; 14 84; 15 19; 16 20; 17 21; 18 22; 19 23; 19 42; 19 59; 20 24; 21 25; 21 82; 22 26; 22 82; 22 84; 23 27; 23 42; 23 61; 23 63; 24 28; 25 29; 25 82; 26 30; 26 61; 26 63; 26 79; 27 31; 28 32; 29 33; 29 75; 29 78; 29 79; 30 34; 30 35; 30 37; 30 63; 30 65; 30 75; 31 35; 32 71; 33 67; 33 68; 33 71; 33 74; 33 75; 33 78; 34 66; 34 67; 34 68; 35 65; 35 66; 35 67; 36 64; 36 65; 36 66; 37 63; 37 64; 37 65; 38 64; 38 65; 38 66; 39 62; 39 63; 39 64; 39 97; 40 62; 40 63; 41 60; 41 61; 41 62; 42 59; 42 60; 42 61; 43 60; 43 61; 43 62; 44 58; 44 59; 44 60; 45 58; 45 59; 46 56; 46 57; 46 58; 46 60; 47 56; 47 57; 48 52; 48 55; 48 56; 48 57; 48 58; 48 60; 48 87; 50 87; 51 55; 51 56; 52 86; 52 87; 53 86; 55 86; 60 87; 62 97; 64 93; 64 96; 64 97; 67 71; 67 72; 67 75; 68 72; 70 74; 71 75; 72 76; 72 90; 73 77; 74 78; 75 79; 76 80; 76 88; 76 90; 77 81; 78 82; 79 85; 79 88; 80 85; 80 88

S6-Cα;6.5 Å SCM (282 contacts)

1 69 ; 1 67 ; 1 68 ; 1 66 ; 1 34 ; 1 35 ; 1 36 ; 2 66 ; 2 67 ; 2 69 ; 2 65 ; 2 92 ; 3 92 ; 3 65 ; 3 66 ; 3 91 ; 3 93 ; 3 38 ; 3 64 ; 3 96 ; 3 97 ; 3 36 ; 4 64 ; 4 65 ; 4 67 ; 4 91 ; 4 92 ; 4 93 ; 4 63 ; 4 90 ; 4 66 ; 4 72 ; 4 69 ; 5 90 ; 5 91 ; 5 63 ; 5 64 ; 5 89 ; 5 62 ; 5 92 ; 5 93 ; 5 94 ; 5 95 ; 5 96 ; 5 97 ; 6 62 ; 6 63 ; 6 65 ; 6 89 ; 6 90 ; 6 61 ; 6 30 ; 6 75 ; 6 79 ; 6 88 ; 6 67 ; 7 88 ; 7 89 ; 7 90 ; 7 61 ; 7 62 ; 7 87 ; 7 43 ; 7 91 ; 8 61 ; 8 63 ; 8 85 ; 8 87 ; 8 88 ; 8 59 ; 8 60 ; 8 26 ; 8 86 ; 8 79 ; 8 89 ; 9 85 ; 9 87 ; 9 59 ; 9 60 ; 9 84 ; 9 86 ; 9 52 ; 9 48 ; 9 57 ; 9 58 ; 10 58 ; 10 59 ; 10 60 ; 10 61 ; 10 84 ; 10 85 ; 10 86 ; 10 57 ; 10 14 ; 10 19 ; 10 22 ; 10 23 ; 10 26 ; 11 84 ; 11 86 ; 11 55 ; 11 57 ; 11 19 ; 11 59 ; 11 83 ; 12 59 ; 12 45 ; 12 57 ; 12 58 ; 12 55 ; 12 86 ; 13 55 ; 13 18 ; 13 86 ; 14 18 ; 14 19 ; 14 22 ; 14 84 ; 15 19 ; 16 20 ; 17 21 ; 18 22 ; 18 84 ; 19 23 ; 19 59 ; 19 42 ; 19 61 ; 20 24 ; 21 25 ; 21 82 ; 22 26 ; 22 63 ; 22 84 ; 22 82 ; 22 85 ; 23 61 ; 23 63 ; 23 27 ; 23 40 ; 23 42 ; 24 28 ; 24 29 ; 25 29 ; 25 79 ; 25 82 ; 26 79 ; 26 30 ; 26 75 ; 26 63 ; 26 85 ; 26 61 ; 27 63 ; 27 37 ; 27 31 ; 27 32 ; 27 40 ; 28 32 ; 29 33 ; 29 75 ; 29 78 ; 29 79 ; 30 35 ; 30 75 ; 30 34 ; 30 67 ; 30 37 ; 30 79 ; 30 63 ; 30 65 ; 31 37 ; 31 35 ; 32 71 ; 33 71 ; 33 67 ; 33 68 ; 33 75 ; 33 78 ; 33 74 ; 34 68 ; 34 66 ; 34 67 ; 35 67 ; 35 66 ; 35 65 ; 35 75 ; 36 65 ; 36 66 ; 36 64 ; 37 64 ; 37 65 ; 37 63 ; 38 64 ; 38 65 ; 38 97 ; 38 66 ; 38 96 ; 39 64 ; 39 62 ; 39 63 ; 39 65 ; 39 97 ; 40 62 ; 40 63 ; 40 61 ; 41 62 ; 41 60 ; 41 61 ; 41 63 ; 41 97 ; 42 60 ; 42 61 ; 42 59 ; 43 59 ; 43 60 ; 43 61 ; 43 62 ; 44 60 ; 44 59 ; 44 58 ; 45 58 ; 45 59 ; 45 57 ; 46 57 ; 46 58 ; 46 60 ; 46 56 ; 47 56 ; 47 57 ; 47 60 ; 48 56 ; 48 57 ; 48 60 ; 48 87 ; 48 52 ; 48 55 ; 48 86 ; 48 58 ; 49 87 ; 50 87 ; 50 55 ; 50 56 ; 51 55 ; 51 56 ; 52 57 ; 52 86 ; 52 87 ; 53 86 ; 55 86 ; 57 86 ; 58 86 ; 60 87 ; 62 97 ; 64 97 ; 64 93 ; 64 95 ; 64 96 ; 65 93 ; 67 72 ; 67 71 ; 67 75 ; 68 72 ; 69 73 ; 70 74 ; 71 75 ; 72 76 ; 72 90 ; 73 77 ; 73 80 ; 74 78 ; 75 79 ; 75 88 ; 75 90 ; 76 88 ; 76 80 ; 76 90 ; 77 81 ; 78 82 ; 79 88 ; 79 85 ; 80 85 ; 80 87 ; 80 88

S6-Cα-Cβ;349 contacts

1 130; 1 132 ; 2 133 ; 2 129 ; 2 71 ; 3 126 ; 3 128 ; 3 130 ; 3 129 ; 4 135 ; 5 181 ; 5 183 ; 5 182 ; 5 180 ; 5 128 ; 5 126 ; 6 71 ; 6 75 ; 6 125 ; 6 129 ; 6 183 ; 6 189 ; 7 125 ; 7 183 ; 7 127 ; 7 122 ; 7 124 ; 7 126 ; 7 178 ; 7 180 ; 7 182 ; 8 181 ; 8 177 ; 8 141 ; 8 135 ; 8 131 ; 8 127 ; 9 182 ; 9 178 ; 9 176 ; 9 122 ; 9 179 ; 9 177 ; 9 174 ; 10 179 ; 10 121 ; 10 183 ; 11 118 ; 11 120 ; 11 122 ; 11 176 ; 11 123 ; 11 121 ; 11 174 ; 12 123 ; 12 155 ; 12 127 ; 12 173 ; 12 177 ; 13 175 ; 13 173 ; 13 118 ; 13 170 ; 13 172 ; 13 174 ; 14 175 ; 14 179 ; 14 121 ; 15 170 ; 15 119 ; 15 117 ; 15 118 ; 15 114 ; 15 116 ; 16 167 ; 16 155 ; 16 123 ; 16 119 ; 16 52 ; 16 173 ; 17 164 ; 17 167 ; 17 169 ; 17 170 ; 17 168 ; 17 166 ; 17 114 ; 18 94 ; 18 171 ; 18 169 ; 18 117 ; 19 169 ; 19 115 ; 19 164 ; 19 168 ; 19 114 ; 19 113 ; 20 167 ; 20 119 ; 20 115 ; 20 44 ; 20 38 ; 21 115 ; 21 169 ; 21 164 ; 21 28 ; 22 28 ; 22 165 ; 22 169 ; 23 88 ; 23 115 ; 23 114 ; 24 88 ; 24 108 ; 24 112 ; 24 169 ; 25 36 ; 26 108 ; 27 36 ; 28 44 ; 28 36 ; 28 38 ; 28 165 ; 29 37 ; 29 36 ; 29 38 ; 30 36 ; 31 38 ; 31 40 ; 31 39 ; 33 40 ; 33 42 ; 33 41 ; 35 42 ; 35 44 ; 35 43 ; 37 46 ; 37 44 ; 37 45 ; 38 115 ; 38 83 ; 38 46 ; 39 46 ; 39 48 ; 39 47 ; 41 49 ; 41 48 ; 41 50 ; 42 161 ; 42 50 ; 43 51 ; 43 50 ; 43 52 ; 44 165 ; 44 161 ; 45 54 ; 45 52 ; 45 53 ; 46 123 ; 46 119 ; 46 83 ; 47 55 ; 47 56 ; 47 54 ; 49 56 ; 49 58 ; 49 57 ; 50 161 ; 51 58 ; 51 60 ; 51 155 ; 51 59 ; 52 123 ; 52 155 ; 52 119 ; 52 60 ; 53 62 ; 53 60 ; 53 61 ; 55 64 ; 55 62 ; 55 63 ; 56 64 ; 57 64 ; 57 65 ; 57 147 ; 57 66 ; 58 153 ; 58 155 ; 58 147 ; 59 68 ; 59 67 ; 59 147 ; 59 69 ; 59 66 ; 60 69 ; 60 73 ; 60 123 ; 60 147 ; 60 127 ; 61 68 ; 63 139 ; 65 131 ; 65 133 ; 65 139 ; 66 145 ; 66 147 ; 66 139 ; 66 153 ; 66 131 ; 67 128 ; 67 131 ; 67 133 ; 67 130 ; 68 131 ; 68 128 ; 68 130 ; 69 127 ; 69 131 ; 70 124 ; 70 127 ; 70 129 ; 70 126 ; 70 128 ; 71 129 ; 72 124 ; 72 127 ; 73 123 ; 73 79 ; 73 127 ; 74 124 ; 74 126 ; 75 125 ; 75 129 ; 76 121 ; 76 124 ; 76 125 ; 76 120 ; 76 123 ; 76 122 ; 77 125 ; 77 121 ; 78 120 ; 78 122 ; 79 123 ; 80 121 ; 80 118 ; 80 120 ; 80 119 ; 80 116 ; 81 121 ; 82 116 ; 82 119 ; 83 119 ; 83 115 ; 84 116 ; 85 121 ; 85 117 ; 85 90 ; 86 115 ; 86 117 ; 86 116 ; 86 113 ; 86 114 ; 87 115 ; 87 114 ; 87 113 ; 88 115 ; 89 112 ; 89 109 ; 89 111 ; 89 113 ; 89 117 ; 90 117 ; 91 109 ; 91 111 ; 91 112 ; 92 112 ; 93 111 ; 93 109 ; 94 171 ; 94 117 ; 94 102 ; 98 171 ; 99 107 ; 99 109 ; 100 110 ; 101 108 ; 101 169 ; 102 171 ; 102 169 ; 104 169 ; 108 169 ; 117 171 ; 125 189 ; 125 183 ; 131 139 ; 131 147 ; 131 141 ; 132 139 ; 132 141 ; 132 140 ; 133 139 ; 134 141 ; 136 145 ; 138 147 ; 138 145 ; 138 146 ; 139 145 ; 140 149 ; 140 147 ; 140 148 ; 141 177 ; 142 150 ; 142 149 ; 142 151 ; 143 151 ; 144 151 ; 144 152 ; 144 153 ; 145 151 ; 146 155 ; 146 154 ; 146 153 ; 147 155 ; 148 157 ; 148 155 ; 148 156 ; 148 173 ; 149 173 ; 149 157 ; 149 177 ; 150 157 ; 150 159 ; 150 158 ; 152 160 ; 152 159 ; 153 159 ; 154 173 ; 154 167 ; 154 161 ; 155 173 ; 156 167 ; 156 173 ; 157 173 ; 160 167 ; 161 167

; 167 173 ; 183 189

M7-Cα;4.5 Å cut-off (253 contacts)

1 16;1 17;1 18;1 19;1 47;1 48;1 49;1 50;2 15;2 16;2 17;2 18;2 19;2 25;2 32;2 46;2 47;2 48;2 49;2 59;3 14;3 15;3 16;3 17;3 45;3 46;3 47;3 48;3 49;4 13;4 14;4 15;4 16;4 17;4 32;4 36;4 44;4 45;4 46;4 47;5 12;5 13;5 14;5 15;5 43;5 44;5 45;5 46;5 47;6 11;6 12;6 13;6 14;6 15;6 36;6 39;6 41;6 42;6 43;6 44;6 45;7 11;7 12;7 13;7 43;7 44;7 45;8 12;8 13;8 39;8 40;8 41;8 42;8 43;9 42;15 32;17 28;17 32;19 24;19 25;21 25;23 27;23 28;24 28;24 29;25 29;25 30;25 48;25 59;25 62;26 30;27 31;28 32;28 33;29 33;29 62;29 63;29 66;30 34;31 35;31 36;32 36;32 37;32 66;33 37;33 38;33 66;33 69;33 70;34 38;34 39;35 39;35 40;36 40;36 41;36 44;36 46;36 66;36 67;36 70;36 72;37 41;37 70;41 72;41 90;42 72;42 88;42 89;42 90;43 72;43 88;43 89;43 90;44 67;44 72;44 86;44 87;44 88;44 89;45 85;45 86;45 87;45 88;46 63;46 66;46 67;46 84;46 85;46 86;46 87;46 88;47 83;47 84;47 85;47 86;48 59;48 63;48 82;48 83;48 84;48 85;49 83;49 84;50 55;50 56;50 59;50 82;50 83;50 84;51 55;51 82;52 56;52 57;52 82;53 57;53 81;53 82;54 58;54 59;55 59;55 60;56 60;56 61;56 79;56 82;56 84;57 61;57 62;57 65;58 62;58 63;58 65;59 63;59 64;59 84;60 64;60 77;60 79;60 84;60 86;61 65;61 66;62 66;62 67;63 67;63 68;63 84;63 86;64 68;64 69;64 75;64 86;65 69;66 70;67 71;67 72;67 75;67 86;67 88;68 72;68 75;71 90;72 88;72 89;72 90;73 89;73 90;74 87;74 88;74 89;74 90;75 86;75 87;75 88;75 89;76 85;76 86;76 87;76 88;76 89;77 84;77 85;77 86;77 87;78 83;78 84;78 85;78 86;79 83;79 84;79 85;80 84;80 85

M7-Cα;SCM (229 contacts)

1 19 ; 1 48 ; 1 49 ; 1 16 ; 1 17 ; 1 18 ; 1 47 ; 2 16 ; 2 17 ; 2 18 ; 2 47 ; 2 48 ; 2 15 ; 2 32 ; 2 25 ; 2 46 ; 2 19 ; 3 16 ; 3 46 ; 3 47 ; 3 48 ; 3 15 ; 3 45 ; 3 49 ; 3 14 ; 4 14 ; 4 15 ; 4 44 ; 4 45 ; 4 46 ; 4 13 ; 4 32 ; 4 36 ; 4 16 ; 4 17 ; 5 14 ; 5 44 ; 5 45 ; 5 46 ; 5 47 ; 5 13 ; 5 43 ; 5 12 ; 6 12 ; 6 13 ; 6 43 ; 6 44 ; 6 11 ; 6 36 ; 6 41 ; 6 15 ; 6 35 ; 6 39 ; 7 43 ; 7 44 ; 7 45 ; 7 11 ; 7 12 ; 7 41 ; 7 42 ; 8 12 ; 8 13 ; 8 42 ; 8 39 ; 8 40 ; 8 41 ; 9 42 ; 15 32 ; 17 28 ; 17 32 ; 17 24 ; 19 24 ; 19 25 ; 21 25 ; 23 27 ; 24 28 ; 24 29 ; 25 29 ; 25 32 ; 25 48 ; 25 59 ; 25 62 ; 26 30 ; 27 31 ; 27 32 ; 28 32 ; 29 33 ; 29 66 ; 29 62 ; 30 34 ; 31 35 ; 31 38 ; 32 36 ; 32 37 ; 32 70 ; 32 66 ; 33 66 ; 33 70 ; 33 37 ; 33 38 ; 33 69 ; 34 38 ; 35 39 ; 35 40 ; 35 41 ; 36 41 ; 36 40 ; 36 72 ; 36 70 ; 36 44 ; 36 46 ; 36 66 ; 36 67 ; 37 70 ; 41 72 ; 41 88 ; 41 90 ; 42 72 ; 42 90 ; 42 88 ; 42 89 ; 43 88 ; 43 89 ; 43 90 ; 44 72 ; 44 87 ; 44 88 ; 44 67 ; 44 86 ; 45 86 ; 45 87 ; 45 85 ; 46 67 ; 46 85 ; 46 86 ; 46 63 ; 46 84 ; 46 66 ; 47 84 ; 47 85 ; 47 83 ; 48 63 ; 48 83 ; 48 84 ; 48 59 ; 48 82 ; 49 82 ; 49 83 ; 50 82 ; 50 83 ; 50 55 ; 50 56 ; 50 59 ; 50 84 ; 51 55 ; 51 56 ; 51 59 ; 51 82 ; 52 82 ; 52 56 ; 53 81 ; 53 82 ; 53 57 ; 54 58 ; 54 59 ; 55 59 ; 56 84 ; 56 60 ; 56 79 ; 56 81 ; 56 82 ; 57 61 ; 57 65 ; 58 65 ; 58 62 ; 59 63 ; 59 86 ; 59 84 ; 60 84 ; 60 86 ; 60 64 ; 60 79 ; 60 77 ; 61 65 ; 61 68 ; 62 66 ; 63 67 ; 63 86 ; 63 68 ; 63 84 ; 64 86 ; 64 68 ; 64 75 ; 65 69 ; 66 70 ; 66 71 ; 66 72 ; 67 72 ; 67 71 ; 67 75 ; 67 88 ; 67 86 ; 68 75 ; 68 72 ; 71 90 ; 72 89 ; 72 90 ; 72 88 ; 73 89 ; 73 90 ; 74 88 ; 74 89 ; 74 87 ; 75 86 ; 75 87 ; 75 88 ; 76 86 ; 76 87 ; 76 85 ; 76 89 ; 77 85 ; 77 86 ; 77 84 ; 78 85 ; 78 83 ; 78 84 ; 79 83 ; 79 84 ; 79 85

Supplementary figures

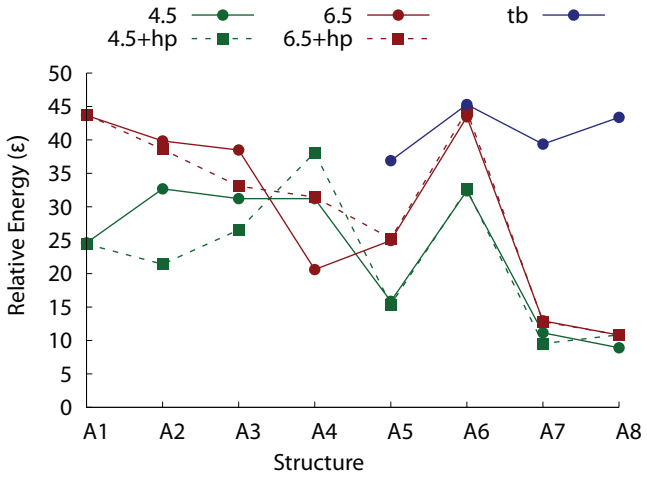

Figure S.1: Energies in kcal/mol (ϵ) of the structures identified as traps relative to the native state. The potentials plotted are the 4.5 Å cut-off based potential with (4.5+hp) and without (4.5) hydrophobic contacts in green, the 6.5 Åcut-off based shadow-contact map potential with (6.5+hp) and without (6.5) hydrophobic contacts in red, and the structures identified from the two-bead potential (tb) are also plotted in blue.

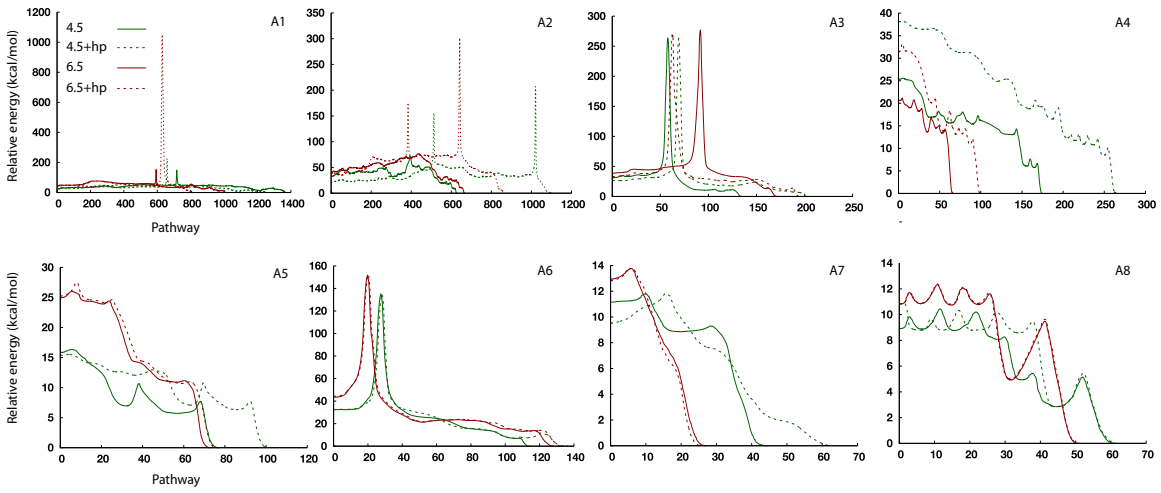

Figure S.2: Potential energies in kcal/mol (ϵ) relative to the native state as a function of the integrated path length along the putative fastest paths for transitioning into the native state from structures A1, A2, A3, A4, A5, A6, A7 and A8 are depicted. The potentials plotted are the 4.5 Å cut-off based potential with (4.5+hp) and without (4.5) hydrophobic contacts and the 6.5 cut-off based shadow-contact map potential with (6.5+hp) and without (6.5) hydrophobic contacts.

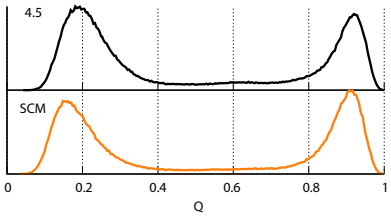

Figure S.3: Normalised population of conformational states (vertical-axis) as a function of fraction of native-contacts, Q (x-axis) derived for the M7 protein from molecular dynamics trajectory at the folding temperature simulations with the 4.5 Å cut-off based potential (4.5) and the 6.5 Å cut-off based shadow-contact map potential (SCM). No intermediate states were found. .